A Neurofunctional Model of Speech Production Including Aspects of Auditory and Audio-Visual Speech Perception

Bernd J. Kröger & Jim Kannampuzha

Department of Phoniatrics, Pedaudiology and Communication Disorders, University Hospital Aachen and Aachen University, Germany

bkroeger@ukaachen.de, jkannampuzha@ukaachen.de

Abstract

A computer-implemented neurofunctional model of speech production is introduced, which is capable of articulating vowels, VC-, and CV-syllables (C = voiced plosives; V = vowels). It will be shown in this paper that this *production* model is capable of simulating basic effects of auditory and audio-visual speech *perception* like (i) categorical perception of consonants and vowels and (ii) the McGurk effect. These typical features of speech perception directly result from the topological ordering of stored speech items at a supra-modal neural level, called a *phonetic map* of this model. This phonetic map is a self-organizing neural map which is trained and structured during early phases of speech acquisition. The neurofunctional model introduced here illustrates the close relationship between speech production and speech perception.

Index Terms: speech production, speech perception, audio-visual speech perception, neurofunctional model, computer simulation, McGurk effect, categorical perception

1. Introduction

Computer-implemented neurofunctional models focusing on the sensorimotor aspects of speech production are rare [1, 2, 3]. The model introduced here is based on the work by Kröger et al. [4, 5]. Its structure is introduced in section 2. Speech knowledge is acquired by training the model (section 3). The main goal of this paper is to demonstrate that this production model is capable of mimicking categorical perception [6] and the McGurk-effect [7] in a straight forward way by using the models feedback sensory paths (see section 4). Experiments supporting this hypothesis are described in section 5 and 6.

2. The Structure of the Model

The organization of the neurofunctional model differentiates cortical and other (i.e. peripheral and subcortical) parts (Fig. 1). The cortical part is subdivided into four parts, i.e. the frontal, the parietal, the occipital, and the temporal lobe. Since speech production or speech movement generation is a sensorimotor process, the model comprises feedforward and feedback control parts [1, 2]. *Feedforward control* starts with the activation of a phonological plan within the phonemic map for the speech item (syllable, word, or utterance) under production. Each *frequent syllable* activates its prestored or prelearned sensory states (auditory, visual, and somatosensory state) and its prestored or prelearned motor plan state via the phonetic map (cf. concept of the mental syllabary [8]). Prestored or prelearned sensory and motor plan states are trained during speech acquisition phases (see section 3). The motor plan of an *infrequent syllable* is generated by the motor planning module. Here, motor plans are assembled by planning information of subsyllabic units (e.g. syllable constituents like onset or rhyme, single speech sounds or single vocal tract actions).

Motor plans represent a high-level motor description of a speech item (syllable or word). A motor plan of a syllable is a score of vocal tract actions (or vocal tract gestures) [9]. Vocal tract actions can be subdivided into four groups [10]. (i) Vocalic vocal tract actions affect the whole vocal tract shape. (ii) Consonantal vocal tract actions realize local labial, apical, or dorsal constrictions or closures gestures. (iii) Velic vocal tract actions distinguish nasal vs. non-nasal and obstruent vs. sonorant sound production and (iv) glottal vocal tract actions distinguish voiced vs. voiceless sound production. On the motor plan level intra- and interaction parameter are determined. Intra-action parameters are duration, action-performing end vocal tract organ, spatial targets for each vocal tract organ, and temporal rapidity for target reaching [10]. Inter-action parameters determine the temporal relationship of different vocal tract actions within a speech item. Consonantal action parameters are, for example, the articulators or vocal tract organs, which perform an opening or closing gesture (e.g. lips, tongue tip, tongue body). Vocalic action targets are described by vocalic values within the high-level vocalic scales low-high, back-front, unrounded-rounded (see section 3).

The time course of positions and velocities for all model articulators are generated by the motor execution module. A three-dimensional articulatory-acoustic model generates the resulting articulatory movement patterns and the resulting acoustic speech signal [10]. The output signals of this model serve as input for somatosensory and auditory *feedback control*. Lower level feedback control comprises somatosensory information which is directly processed by parts of the motor execution module. Higher level feedback control comprises somatosensory (tactile and proprioceptive) and auditory information. Auditory and somatosensory states of the currently produced speech item are forwarded towards the auditory and somatosensory processing units. Here an auditory and somatosensory error signal is calculated by comparing the current sensory state of a speech item (activated within primary sensory maps) with its prestored sensory state (activated within the sensory state maps) [1, 2]. The prestored sensory state is already activated during forward control via the phonetic map (see above). This error signal can be used for correcting the motor state for the speech items under production.

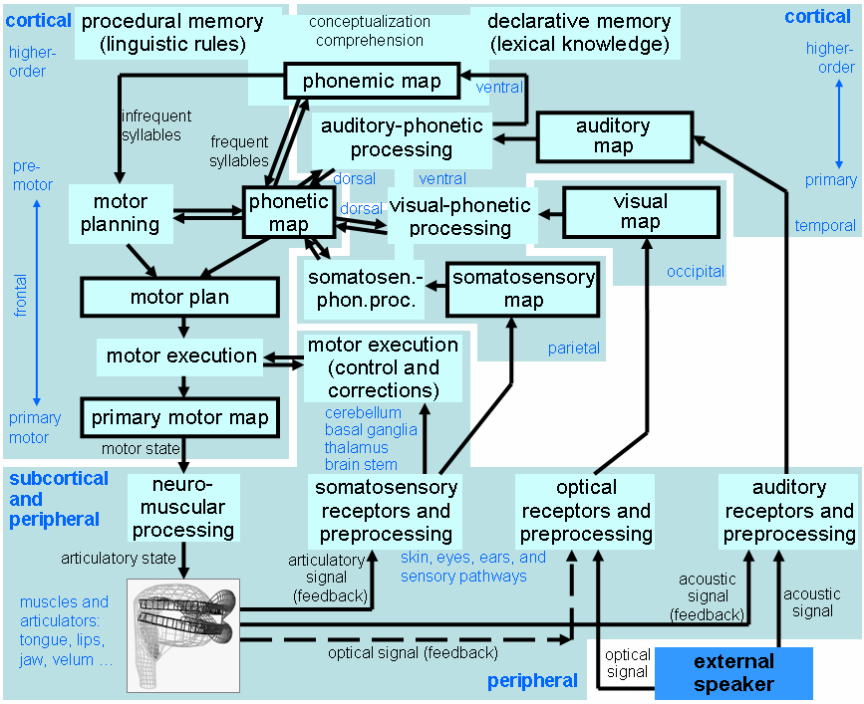

Figure 1: *Neurofunctional model of speech production. Boxes with black outline represent neural maps, arrows indicate processing paths or neural mappings. Boxes without black outline indicate processing modules (to be specified in detail).*

The auditory feedback pathway described above is also used as the auditory path for processing *external* acoustic speech signals (i.e. speech signals produced by external speakers). Since subjects normally have no access to tactile and proprioceptive (i.e. somatosensory) signals produced by external speakers, these somatosensory feedback pathways are used for feedback control exclusively. In the case of visual signals the situation is vice versa. In this case the subject normally has no access to the visual result of his/her own productions. Thus the visual pathway is used mainly for processing optical signals generated by external speakers, i.e. optical signals produced in the case of face-to-face communication in parallel to the acoustic signal of an external speaker. Only in the rare case of speaking and monitoring the subject's own face in a mirror (which sometimes is used as a bio-feedback technique in speech therapy), visual signals are used for feedback control.

The somatosensory, auditory, and visual pathways (i.e. sensory pathways) within our model lead to neural activation patterns which represent the sensory state of the currently produced speech item. On the level of the auditory state map, the neural activation pattern represents the bark scaled formant trajectories of F1, F2, and F3 of the currently produced speech item (e.g. syllable). On the level of the somatosensory state map, the neural activation pattern represents the time course of proprioceptive and tactile parameters. Proprioceptive state parameters in our model coincide with motor plan parameters: the actual displacement and actual movement velocity for the parameters jaw vertical position (jaw angle), tongue body vertical position (or angle) and horizontal position, tongue tip vertical position (or angle) and horizontal position, lip opening distance, lip protrusion, and in addition two hyoid parameters and one laryngeal parameter. Tactile parameters in our model are contact area of articulators (lips, tongue tip, tongue body) with vocal tract walls (alveolar ridge, postalveolar palatal, velar, and pharyngeal region). On the level of the visual state map, the neural activation pattern in our model represents the time course of optical facial parameters of the mouth region, i.e. lip opening distance and lip protrusion. These two parameters are also part of the proprioceptive state description.

In the case of feedback perception, the speaker's sensory signals are used for correcting feedforward motor signals via the phonetic map (dorsal pathway or dorsal stream). In the case of perceiving external speakers, the auditory and/or visual signals are processed via two pathways. These external signals (i) directly activate lexical items via the auditory-to-meaning path or ventral stream [11] (comprehension) and/ or these signals (ii) activate motor states via the phonetic map via the auditory-motor path or dorsal stream [11] (segmental perception). The interconnection of the higher level sensory (auditory, somatosensory, and visual) processing modules (see Fig. 1) indicates that auditory, visual, and somatosensory states can be interpreted in a hyper-, multi-, or supramodal representation or "hypermodal percept" (for a complete survey and discussion of the supra-modal neural representation hypothesis see [12]).

3. Simulation of Speech Acquisition

The *structure* of the neurofunctional model is described above. The speech *knowledge*, i.e., how to produce a certain sound or syllable, is gained during learning or training phases. These training phases simulate the early phases of speech acquisition [13]. In the model described here, the sensorimotor speech knowledge is mainly stored in a self-organizing *phonetic map* (Fig. 1), i.e. within the *neural link weights of the neural mappings* between this phonetic map and the phonemic map, the sensory state maps, and the motor plan map [5]. Currently the computer-implemented version of our neuronfunctional model is capable of processing V-syllables for a 5 vowel system (/i/, /e/, /a/, /o/, and /u/) and CV- and VC-syllables for this 5 vowel system in any combination with one of three voiced plosives /b/, /d/, and /g/. Training procedures in general can be separated in two consecutive phases.

During the *babbling phase* the model learns to relate sensory to motor states for pre-linguistic proto-vocalic and proto-consonantal articulations. For each type of (proto-) syllable (V, CV, VC, …) a self-organizing map arises as a part of or as a submap of the phonetic map. Each neuron within a phonetic submap comprises related motor plan and sensory state information, i.e. link weight values for motor and sensory parameters (see Fig. 2; the list of sensory parameters is given above).

After this babbling phase, the model is capable of imitating motor states just by activating sensory (e.g. auditory or audio-visual) states. Thus, during the following *imitation phase* the model perceives and reproduces (i.e. imitates) externally produced acoustic speech items of a specific target language. After the imitation phase the model has acquired these language-specific speech items (syllables or words) and

is capable of producing these items if this item is activated on the level of the phonemic map.

At the level of the phonetic map, the imitation training (i) leads to a further refinement of the motor plan and sensory link weight values connecting the phonetic map with the motor state and sensory state maps and (ii) leads to the formation of phonemic link weight values, i.e., leads to the formation of the neural connections between the phonetic and phonemic map. Certain neurons within the phonetic map then represent an amount of sensory and motor plan states for phonemic states. These neurons are indicated by a solid or dashed outline in Fig. 2 for /b/, /d/, and /g/ in the case of a phonetic CV-map.

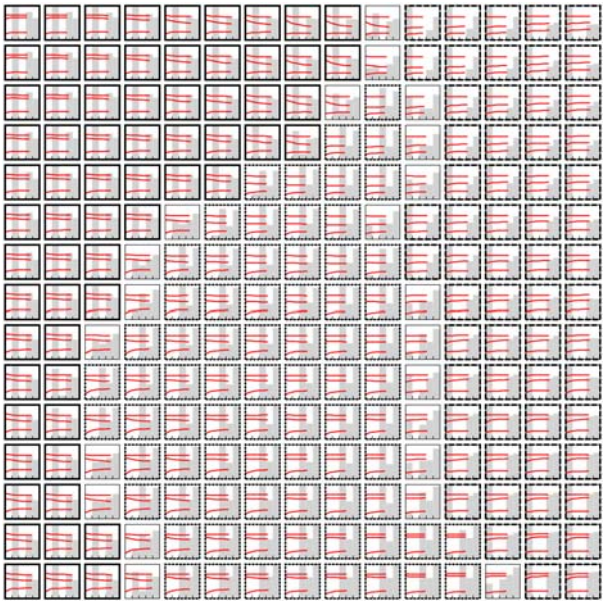

Figure 2: *Self-organizing phonetic map (15x15 neurons) for virtual listener 4 (see text) for CV-syllables after babbling and imitation training (C = voices plosives). Bars within each neuron square: motor plan parameters; first three bars: vocal tract organ which performs the closing gesture (labial, apical, dorsal); two last columns: back-front value (forth column) and low-high value (fifth column) of the vowel within the CV-syllable. Horizontal trajectories within each neuron square: auditory state parameters: bark scaled F1, F2, and F3 formant transitions. Somatosensory and visual link weights are not shown. The outlines of the neuron boxes give the phonemic state. Dashed: /b/; solid: /d/; dotted: /g/. This is the map of model instance 4.*

If a neuron representing a speech item is activated within the phonemic map, it directly coactivates a neuron (or an ensemble of neighboring neurons) in the phonetic map and in addition directly coactivates the sensory states and the motor plan of this speech item via the phonetic map. The activation of the motor plan state then activates and starts the motor execution of the speech item.

But why is the phonetic map introduced in our model? It could be argued, that there are direct neural connections between the phonemic map, the motor plan map, and the sensory state maps (cf. [1, 2]). The answer is that the neural connection between these maps is modeled in our approach by using *self-organizing maps* [14]. The phonetic map is the central self-organizing map, summarizing the knowledge for the correct neural connection of speech states (i.e. of phonemic, motor plan, and sensory states) for a speech item under training. Thus, the self-organizing map can be seen on one hand just as a part of the mapping between phonemic, motor plan, and sensory states. But, as will be shown below, on the other hand the ordering of phonetic states, as occurs in this map, directly reflects phonetic knowledge. For example, vocalic speech items within these maps are ordered with respect to phonetic parameters like low-high, back-front [5], or CV- or VC-items are ordered with respect to place of articulation ([5] and see Fig. 2).

4. Using the Model for Speech Perception

Beside the feedforward production pathway (phonemic state → phonetic state → motor plan state; Fig. 1) the model also is capable of identifying the phonemic state of a speech item after imitation training by using the dorsal perception pathway (auditory or audio-visual state → phonetic state → phonemic state). This dorsal perception pathway [11] activates prelearned production knowledge: If the auditory state of a frequent syllable is activated, the phonemic state of this syllable is coactivated via this dorsal pathway. Thus, the phonetic map not only plays an important role in speech production but also in speech perception.

Beside speech production also speech perception, especially the identification and discrimination of speech sounds, can be modeled easily on the level of the phonetic map in our approach. Perceptual *identification* is modeled as follows: An externally produced formant pattern activates that neuron within the phonetic map, which represents the most similar formant pattern (formant patterns represented by phonetic map neurons are displayed in Fig. 2 for the case of CV-syllables). This maximal activated neuron then coactivates a syllable neuron (for example in the case of CV-syllabels in our model the /ba/, /da/, or /ga/, /bi/, /di/, /gi/,…. neuron) within the phonemic map. Perceptual *discrimination* of two speech items is modeled by calculating the city-block distance between the maximal activated neurons for both speech stimuli on the level of phonetic map. The smaller the distance of two speech items, the worse is the perceptual discriminatory power.

It will be shown in the following chapters of this paper that this modeling of identification and discrimination on the level of the phonetic map leads to a correct prediction of well-known speech perception effects like categorical perception or the McGurk-effect.

5. Simulation of Categorical Perception

Purpose: A typical effect of speech perception is *categorical perception*. Categorical perception typically occurs for consonants, while the perception of vowels is more continuous or less categorical [6]. We hypothesize that this perception effect can be simulated in our production model after babbling and imitation training. **Method:** 20 different instances of the model were generated by performing babbling and imitation training. The training of these instances of the model was done using different initial link weight values and different ordering of training items [4]. Thus 20 "virtual" subjects, called "virtual speakers" or "virtual listeners" were generated, leading to 20 different V- and VC-phonetic maps. In order to demonstrate categorical perception, in addition a vocalic and a consonantal acoustic stimulus continuum was generated, comprising 13 stimuli which cover the /i/-/e/-/a/-range for vowels and 13 stimuli which cover the /ba/-/da/-/ga/-range for

consonants. The consonantal stimulus continuum is displayed in Fig. 3.

Figure 3: *Bark scaled formant trajectories of the /ba/-/da/-/ga/-stimulus-continuum (stim. 1 to stim. 13).*

Results: *Identification scores* (Fig. 4 and 5, thin black lines) were calculated by accumulating the individual vowel or consonant identifications for each stimulus. This is done by identifying the maximal activated neuron of the phonetic map for each stimulus and for each virtual listener and by identifying the strongest phonemic link weight value for this neuron. *Discrimination scores* (Fig. 4 and 5, thick black line) were calculated by accumulating the individual distances of stimulus-pairs on the level of the phonetic map. Only stimulus-pairs with a constant acoustic stimulus distance were chosen; here: stim. 1 and stim. 3, stim. 2 and stim. 4, … , stim. 11 and stim. 13, see Fig. 3 for the consonantal stimuli). Beside measured discrimination (thick black line) also calculated discrimination, i.e. discrimination calculated from measured identification scores is displayed in Fig. 4 and Fig. 5 (for a definition of calculated discrimination see [15]).

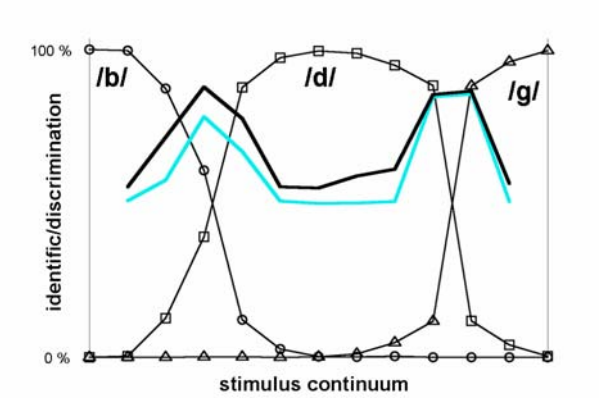

Figure 4: *Discrimination scores (measured and calculated) and identification score (measured) for an acoustic /ba/-/da/-/ga/-continuum (continuum of voiced plosives).*

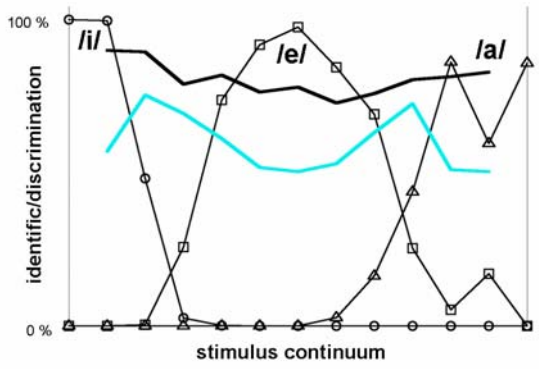

Figure 5: *Discrimination scores (measured and calculated) and identification score (measured) for an acoustic /i/-/e/-/a/-continuum.*

Discussion: Typical identification and discrimination scores for strong categorical perception occur for voiced plosives (Fig. 4), while the effect of categorical perception is less pronounced in the case of vowels (Fig. 5). In the case of vowels, in addition the occurring gap between measured and calculated discrimination indicates that vocalic sounds comprise more acoustic information than just distinctive speech features.

6. Simulation of the McGurk Effect

Purpose: The McGurk effect is a well-known effect of audiovisual speech perception [7]: A [gaga] video (optical signal), which is dubbed with a [baba] audio signal (acoustic signal), is perceived by most subjects as [dada]. We hypothesize that this effect can be simulated using our neurofunctional model of speech production after babbling and imitation training.
Method: 20 instances of the model were trained using different initial link weight values and different ordering of training items. In addition to the auditory processing, the somatosensory parameters for lip opening distance and lip rounding are doubled now for being able to model the visual processing (Fig. 1). Lip data are added here for babbling and imitation training. If a model instance is exposed to the McGurk-stimulus, the maximal activated neuron of the phonetic map must be calculated as is done in other identification tasks. But due to the neural mapping from the visual state map to the phonetic map, all those neurons of the phonetic map are in an *inhibitory state* (i.e., can not be activated), which represent a weak, medium, or strong lip closure, since the visual McGurk-signal indicates *no* lip closure. Thus in the case of the McGurk stimulus that neuron of the phonetic map is maximal activated (i.e. is the winner neuron), which is not visually inhibited by the stimulus and which indicates the most similar formant pattern with respect to the auditory part of the McGurk stimulus. This neuron is marked as the *non-crossed bold outlined neuron* in Fig. 6-8. The *crossed bold outlined neuron* in these figures represents the maximal activated neuron in the case of the pure auditory [baba] stimulus, which is part of the McGurk stimulus (condition: no visual signal is applied). The link weight values for the motor plan parameters "closure-performing articulator" of all neurons of the phonetic map – which can be interpreted as the percentages of labial, apical, or dorsal activation – are indicated by the first three grey bars for each neuron in Fig. 6 to 8. It can be assumed that these link weight values are linked with the phonemic activation for /b/, /d/, or /g/. Furthermore we assume that (i) the virtual listener always perceives the phoneme which is linked with that neuron of the phonetic map, which is maximal activated by the McGurk stimulus, i.e., maximum motor plan link weight value of the parameters labial, apical, or dorsal), e.g. /d/ for virtual listener 4 and 2 (Fig. 6 and 7) and /g/ for listener 3 (Fig. 8) (*few exposures assumption*), or that (ii) the probability for perceiving a distinct phoneme /b/, /d/, or /g/ is proportional to all motor plan link weight values for labial, apical, and dorsal for the same neuron within the phonetic map, which is maximally activated by the McGurk-Stimulus (*many exposures assumption*).

Results: The perceptual identification scores for the McGurk stimulus are given in Table 1 for the *many exposures assumption*. Here 100 identification tasks are assumed for each virtual listener and each stimulus. From the results of our simulation experiment given in Fig. 6 to 8 and given in Table 1, it can be seen that there exist three different *types* of virtual listeners with respect to the perceptual neural processing of the McGurk stimulus. *Listener type 1* always perceives a /d/ (virtual listener 4, 7, 8, 11, and 17 in Table 1; see also Fig. 6). For this type of listener, the maximal

activated neuron (winner neuron) within the phonetic map is spatially separated from the region of inhibited "visual labial" neurons and thus also the pure auditory [baba] winner neuron is spatially separated within the phonetic map from the McGurk winner neuron. *Listener type 2* mainly perceives a /d/, but this type of listener also perceives /b/ in some cases, if he is exposed to the McGurk stimulus many times (virtual listener 1, 2, 5, 6, 9, 10, 12, 15, 19, 20 in Table 1; see also Fig. 7). For this type of listener, the maximal activated neuron (winner neuron) within the phonetic map is a neighboring neuron with respect to the region of inhibited "visual labial" neurons, but the distance between the McGurk winner neuron and the pure auditory [baba] winner neuron is less than for listener type 1. *Listener type 3* mainly perceives /g/, but this type of listener also perceives /b/ in some cases, if he is exposed to the McGurk stimulus many times (virtual listener 3, 13, 14, 16, 18 in Table 1; see also Fig. 8). For this type of listener, the maximal activated neuron (winner neuron) within the phonetic map indicates the same spatial relations as occur for listener type 2: close neighboring to the inhibited "visual labial" region, but medium distance to the neuron, which is maximal activated in the case of the auditory [baba] winner neuron.

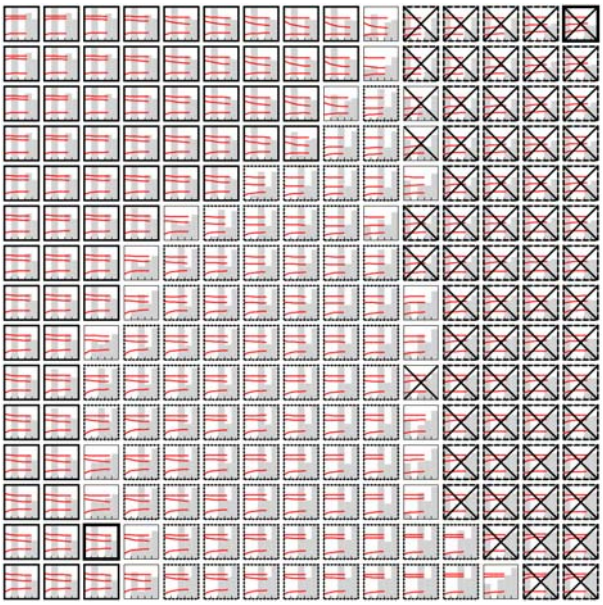

Figure 6: *Self-organizing phonetic map (15x15 neurons) for CV-syllables after babbling and imitation training (C = voices plosives) for virtual listener 4 (see also Fig. 2). Two neurons outlined by bold lines indicate the winner neurons in the case of auditory-only (bold outlined and crossed neuron) and audio-visual perception (bold outlined and non-crossed neuron).*

In addition three levels of inhibitory strength within the neuroperceptual part of our neurofunctional production model were tested: weak, medium, and strong. In the case of the *strong inhibition assumption*, neurons within the phonetic map were completely inhibited in the case of exposure of the model by the McGurk stimulus, if these neurons represent just a part of a labial closing movement. No complete labial closure needs to be represented by these inhibited neurons. This strong inhibition assumption is used in our simulations presented above (Fig. 6 to 8 and Table 1). In the case of the *weak inhibition assumption*, only those neurons are completely inhibited, which represent a strong closure, i.e. a complete closure for a distinct time period. In the case of the *medium inhibition assumption*, only those neurons are completely inhibited, which represent a labial closing movement towards a weak closure, i.e. lips just come in contact. Perceptual results for the exposure of the McGurk stimulus to the model are given for all assumptions (strong, medium, and weak inhibition assumption; few and many exposures assumption) in Table 2. These results indicate that the strong inhibition assumption is the most realistic assumption for simulating the McGurk effect in our neural model of speech production (cf. perception rates in [7]).

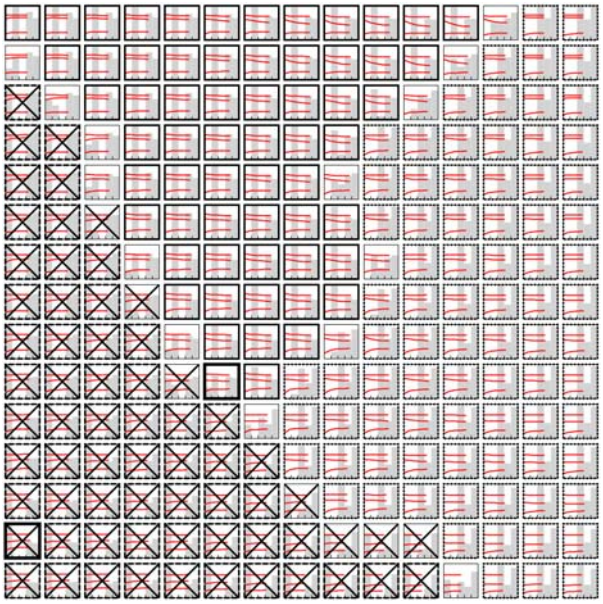

Figure 7: *Same as Fig. 6 but for a different instance of the model (virtual listener 2).*

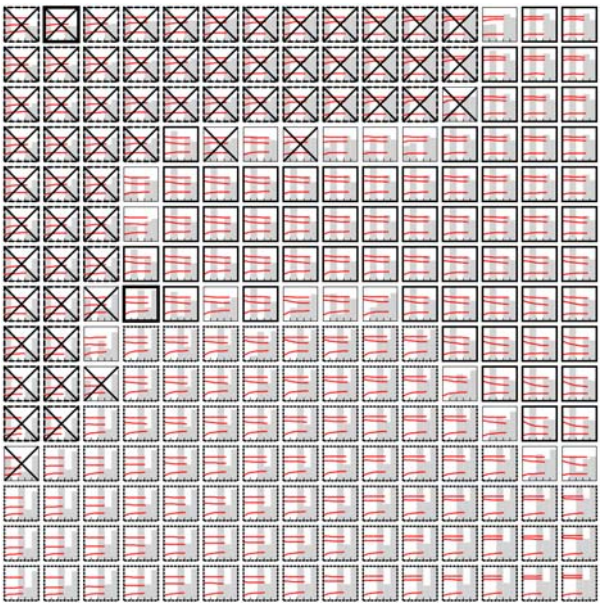

Figure 8: *Same as Fig. 6 but for a different instance of the model (virtual listener 3).*

Discussion: This experiment indicates that the McGurk effect can be simulated easily using the perception paths of our production model. Furthermore different types of virtual listeners can be identified in the case of the McGurk effect. This may

reflect that also in reality different types of listeners occur. Some subjects clearly perceive a /d/ while other subjects get unclear percepts by being exposed to the McGurk stimulus.

Table 1: *Perceptual results for the McGurk-[baba]$_{acoustic}$-[gaga]$_{optical}$-stimulus for all 20 instances of the model (virtual listeners). Percentage of /b/, /d/, or /g/ perception in the case of 100 identification tasks per virtual listener and in the case of the strong inhibition assumption. The highest activation rate is given in bold letters for each listener.*

Listener	/b/	/d/	/g/
1	0.20	**0.80**	0.00
2	0.33	**0.67**	0.00
3	0.00	0.09	**0.92**
4	0.00	**1.00**	0.00
5	0.34	**0.66**	0.00
6	0.27	**0.73**	0.00
7	0.00	**1.00**	0.00
8	0.00	**1.00**	0.00
9	0.04	**0.96**	0.00
10	0.34	**0.66**	0.00
11	0.00	**1.00**	0.00
12	0.15	**0.85**	0.00
13	0.19	0.00	**0.81**
14	0.08	0.00	**0.92**
15	0.34	**0.66**	0.00
16	0.21	0.00	**0.79**
17	0.00	**1.00**	0.00
18	0.29	0.00	**0.71**
19	0.26	**0.74**	0.00
20	0.39	**0.61**	0.00

Table 2: *Perceptual results for the McGurk-[baba]$_{acoustic}$-[gaga]$_{optical}$-stimulus exposed to 20 instances of the model (virtual listeners). Percentage of /b/, /d/, or /g/ perception in different cases. Columns: Amount of exposure per stimulus and per virtual listener; rows: strong, medium, or weak inhibition assumption (see text).*

	one exposure			**100 exposures**		
	/b/	**/d/**	**/g/**	**/b/**	**/d/**	**/g/**
strong	0.00	0.75	0.25	0.17	0.62	0.21
medium	0.25	0.55	0.20	0.32	0.50	0.18
Weak	0.50	0.30	0.20	0.43	0.39	0.18

7. General Discussion and Conclusions

The perception experiments performed in this study indicate that the computational neurofunctional model of speech production introduced here is capable of demonstrating characteristic features of speech perception after babbling and imitation training. These features mainly result from the *topological organization of speech items on the level of the phonetic map,* which develops during model training phases (i.e. during speech acquisition). Thus the results of the simulation experiments outlined in this study indicate the *close relationship of speech production and speech perception* as has been postulated for example by the motor theory of speech perception [16]. It will be shown in further studies that a complete neural model for auditory and audiovisual speech perception (as is discussed for example in [11] and in [17]) can be integrated easily and in a straight forward way into our production model.

8. Acknowledgements

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

9. References

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

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

[3] Bailly G (1997) Learning to speak: sensory-motor control of speech movements. Speech Communication 22: 251-267

[4] Kröger BJ, Birkholz P, Kannampuzha J, Neuschaefer-Rube C (2007) Modeling the perceptual magnet effect and categorical perception using self-organizing neural networks. Proceedings of the 16th International Congress of Phonetic Sciences (Saarbrücken, Germany) 789-792

[5] Kröger BJ, Lowit A, Schnitker R (2008) The organization of a neurocomputational control model for articulatory speech synthesis. In: Esposito A, Bourbakis N, Avouris N, Hatzilygeroudis I (eds.) Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction. (Springer, Berlin) 126-141

[6] Eimas PD (1963) The relation between identification and discrimination along speech and non-speech continua. Language and Speech 6: 206-217

[7] McGurk H, MacDonald J (1976) Hearing lips and seeing voices, Nature 264: 746-748

[8] Levelt WJM, Roelofs A, Meyer A (1999) A theory of lexical access in speech production. Behavioral and Brain Sciences 22: 1-75

[9] Goldstein L, Byrd D, Saltzman E (2006) The role of vocal tract action units in understanding the evolution of phonology. In: Arbib MA (ed.) Action to Language via the Mirror Neuron System. (Cambridge University Press, Cambridge) 215-249

[10] 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, Berlin) 174-189

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

[12] Schmid G, Ziegler W (2006) Audio-visual matching of speech and non-speech oral gestures in patients with aphasia and apraxia of speech. Neuopsychologia 44: 546-555

[13] Oller DK, Eilers RE, Neal AR, Schwartz HK (1999) Precursors to speech in infancy: the prediction of speech and language disorders. Journal of Communication Disorders 32: 223-245

[14] Kohonen T (2001) Self-organizing maps (Springer, Berlin NewYork)

[15] Liberman AM, Harris KS, Hoffman HS, Griffith BC (1957) The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54: 358-368

[16] Liberman AM, Mattingly IG (1985) The motor theory of speech perception revised. Cognition 21: 1-36

[17] Ojanen V (2005) Neurocognitive mechanisms of audiovisual speech perception. Doctoral dissertation. Helsinki University of Technology, Laboratory of Computational Engineering, Technical report Nr. B49 (Helsinki, Finland).