Categorical Perception of Consonants and Vowels: Evidence from a Neurophonetic Model of Speech Production and Perception

Bernd J. Kröger, Peter Birkholz, Jim Kannampuzha,
and Christiane Neuschaefer-Rube

Department of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and RWTH Aachen University, Aachen, Germany
{bkroeger,pbirkholz,jkannampuzha,cneuschaefer}@ukaachen.de

Abstract. While the behavioral side of categorical perception in speech is already well investigated, little is known concerning its underlying neural mechanisms. In this study, a computer-implemented neurophonetic model of speech production and perception is used in order to elucidate the functional neural mechanisms responsible for categorical perception. 20 instances of the model ("virtual listeners/speakers") underwent a speech acquisition training procedure and then performed behavioral tests, i.e. identification and discrimination experiments based on vocalic and CV-syllabic speech stimuli. These virtual listeners showed the expected behavioral results. The inspection of the neural organization of virtual listeners indicated clustering in the case of categorical perception and no clustering in the case of non-categorical (continuous) perception for neurons representing the stimuli. These results highlight a possible neural organization underlying categorical and continuous perception.

Keywords: speech perception, categorical perception, identification, discrimination, neural model of speech production.

1 Introduction

Categorical perception is an important feature of speech, needed for successfully differentiating and identifying sounds, syllables, or words. Categorical speech perception enables humans to map different realizations of one speech sound into one category. This is important in order to achieve a robust discrimination of different speech items, even if these items are realized by different speakers or by different articulations of the same speaker. A quantitative definition of categorical perception, based on identification and discrimination experiments, was given decades ago [1]. Based on this definition it was shown that a consonantal stimulus continuum covering the /ba/-/da/-/ga/-range is perceived more categorically than a vocalic stimulus continuum covering the /i/-/e/-/a/-range [2]. It is unclear whether pure continuous perception occurs at all since even non-speech acoustic stimulus continua like e.g. single tone stimulus continua (pitch perception) indicate a tendency to categorical perception

A. Esposito et al. (Eds.): COST 2102 Int. Training School 2010, LNCS 6456, pp. 354–361, 2011.

(for a discussion of non-speech continuous or categorical perception see [3], [4], [5], and [6]). Currently, the interest in categorical versus continuous perception again increases, because the underlying neural mechanisms of continuous and categorical perception are not yet resolved (see [7] and [8]) and because categorical perception algorithms are needed for the construction of artificial agents [9].

On the basis of computer simulation experiments a neural mechanism will be identified in this study, which could be responsible for categorical vs. continuous perception of acoustic stimulus continua. These experiments are based on a neurophonetic model of speech production and speech perception, which is capable to reproduce the quantitative results of behavioral identification and discrimination experiments occurring for consonantal /ba/-/da/-/ga/- and vocalic /i/-/e/-/a/-stimulus continua [10].

2 The Neurophonetic Model

Our neurophonetic model (Fig. 1) can be divided in a motor feed forward part (from phonemic map via phonetic map or motor planning module to articulatory states) and a sensory feedback part (from sensory preprocessing via sensory-phonetic processing to syllabic auditory map SAM and to syllabic somatosensory map SSM).

In the case of the production (motor feed forward part) of frequent syllables (i.e. already acquired syllables), each syllable is coded by one model neuron on the level of the phonemic map. Activation of that phonemic state (e.g. /ba/) leads to a co-activation of one or more neurons within the phonetic map further co-activating a motor plan, syllabic auditory and somatosensory state for that syllable. At the motor plan level, a gesture score is activated for that syllable, i.e. a high level motor description of all vocal tract actions and their temporal coordination needed for producing that syllable; i.e. consonantal bilabial closing action, vocalic tongue lowering action, and glottal closing action for phonation and the temporal coordination of these vocal tract actions in the case of /ba/ [11]. This motor plan state leads to an activation of specific primary motor states (i.e. articulator positions and movements; cp. [12]) for each time instant during the execution of the syllable.

Within the sensory feedback part, a current somatosensory and auditory state (sensory state) is passed from the periphery (receptor neurons and preprocessing modules) towards the somatosensory and auditoy map for each time instant. The sensory state is stored at the level of the syllabic auditory and syllabic somatosensory map (SAM and SSM as part of the working or short-term memory). Auditory preprocessing is implemented currently by extracting the formant frequency of the first three formants with a time step of 10 ms. Thus the trajectories of the first three formants F1, F2, and F3 over time are stored for a whole syllable within the SAM (syllable formant pattern). In the same way tactile information (i.e. contact area of lips, hard and soft palate) and somatosensory information (i.e. current position and movement velocity of upper and lower lips, tongue tip, tongue body and lower jaw) is updated each 10 ms and stored temporarily in the SSM for produced and currently perceived syllables.

In order to provide the model with speech knowledge, a babbling training and afterwards an imitation training is performed. During babbling training syllabic sensory

states related to syllabic motor states were generated for an amount of 2158 random proto-V and 2079 random proto-CV training items (V = vocalic; C = consonant). 1000 training steps were calculated per training item. The babbling training leads to an adjustment of neural link weights between the phonetic map (i.e. a a self-organizing map or SOM) and the motor plan, syllabic auditory, and syllabic somatosensory map [13]. Currently, two separate 15x15 SOM's were trained for V- and CV-items, i.e. the V-part and the CV-part of the phonetic map. After language-independent babbling training the same SOM's and their neural association to motor map, sensory maps and now also to the phonemic map were further trained now by imitating language specific 6125 V- and 6255 CV-stimuli. This was done for a "model language" comprising five vowels (/i/, /e/, /a/, /o/, /u/), three consonants (/b/, /d/, /g/), and all 15 CV-combinations (C = consonant, V = vowel) of these speech sounds. Again, 1000 training steps were calculated per training item. After babbling and imitation training the model was capable to produce and to perceive these 5 trained vowels and the 15 trained syllables.

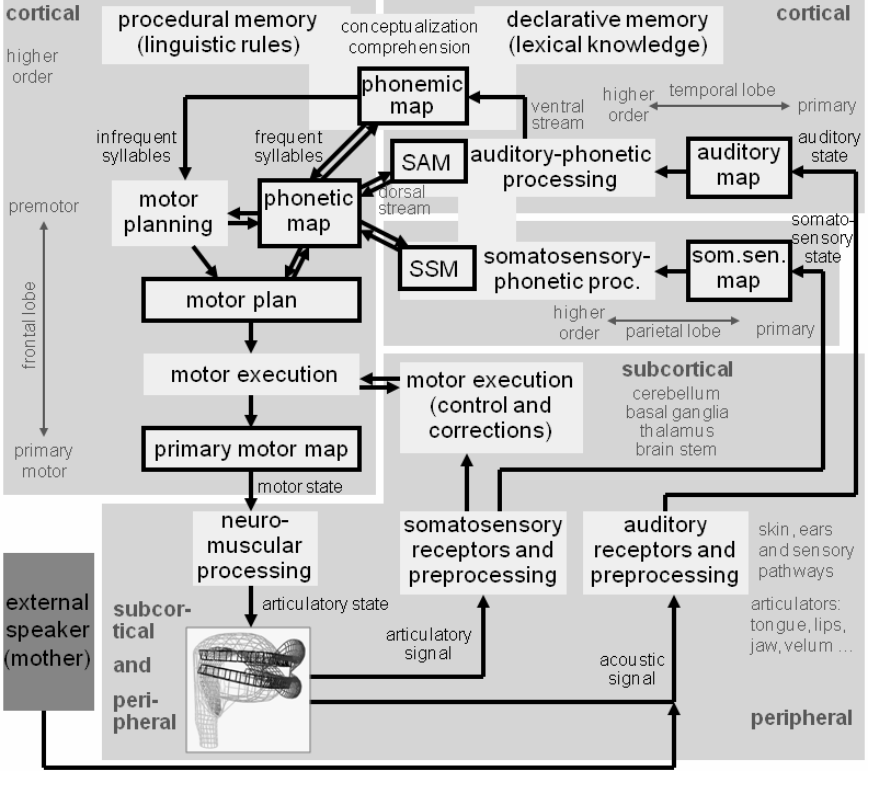

Fig. 1. Organization of the neurophonetic model of speech production and perception. Framed boxes indicate neural maps, arrows indicate neural mappings or processing paths, non-framed boxes indicate processing modules (see also text).

3 Method

20 instances of the model (i.e. "virtual listeners/speakers") were trained using a different initialization of the phonetic to phonemic, sensory, and motor plan map. The knowledge which is acquired for each instance of the model during the babbling and imitation phase (i.e. during the earliest phases of speech acquisition) is stored within the bidirectional neural associations of phonetic to other maps as described above. Thus a phonetic state, which is represented by a neuron within the phonetic map, represents (i) a realization of a phonemic state /V/ or /CV/, (ii) a motor plan state, (iii) an auditory state, and (iv) a somatosensory state. Consequently, the activation of a phonetic state via the phonemic map – as it occurs during production – means that the speaker knows, how to produce that speech item (motor plan), what that speech item sounds like (auditory map), and what the production of that speech item "feels" like (somatosensory map). Moreover, the activation of a phonetic state via the auditory path (i.e. via activation of a syllabic auditory state) – as it occurs during perception – means that the listener is capable to identify its phonemic state via most strongly co-activated neuron within the phonemic map. We tried to visualize this complex information associated with each neuron of the phonetic map in Fig. 2 for the V-part and in Fig. 3 for the CV-part of the phonetic map. 15x15 maps were chosen for the V- as well as for the CV-part (C=/b, d, g/) of the phonetic map (see section 4 of this paper).

After speech acquisition training (i.e. babbling and imitation), the 20 instances of the model as virtual listeners can perform identification and discrimination experiments for a vocalic /i/-/e/-/a/- stimulus continuum (13 stimuli) and for a consonantal /ba/-/da/-/ga/-stimulus continuum (13 stimuli as well, see Kröger et al. 2009). Listening to each of these acoustic stimuli leads to a specific neural activation pattern at the syllabic auditory state level (SAM, Fig. 1) and subsequently to a specific co-activation at the level of the phonetic map. Thus within the phonetic map for each of the 13 V- and 13 CV-stimuli one (winner) neuron can be identified, which exhibits the highest activation and thus represents the phonetic state of that stimulus. These "stimulus neurons" are indicted in Fig. 2 and 3 by bold outlined boxes for the V- and the CV-part of the phonetic map for a sample virtual listener (i.e. model instance 11).

4 Results

Fig. 2 and 3 display the V- and the CV-part of a phonetic map including the visualization of link weights (i.e. strength of connection of neurons) between phonetic and auditory and phonetic and motor plan map for a typical virtual listener (listener 11). A first inspection of this phonetic map indicates that the winner neurons representing the phonetic states of the 13 V-stimuli (also called V-stimulus neurons) are distributed nearly continuously while the CV-stimulus neurons are clustered into 3 groups. And it can be seen that this clustering is related to phoneme regions, i.e. to the /ba/-, /da/-, and /ga/-region (Fig. 3).

We tested the hypothesis that the tendency for clustering of states is higher for the CV- than for the V-case by analyzing the location of all test stimulus neurons within the phonetic maps for all 20 instances of the model. For this analysis two criteria were set for identifying a stimulus neuron cluster: (i) Neighboring neurons within a cluster

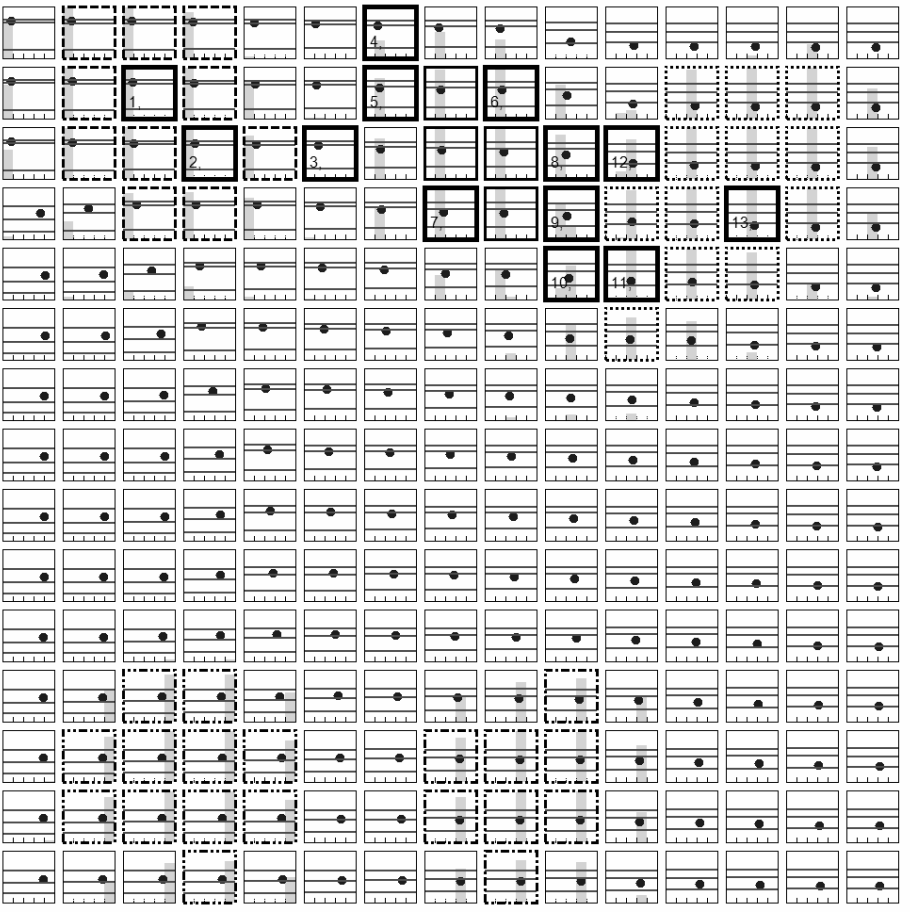

Fig. 2. Phonemic, motor, and sensory states represented by each neuron of the phonetic map for vowels (model instance 11). Each box represents a neuron of the phonetic map for vowels. Light grey bars represent the degree of activation of a phonemic state (from left to right: degree of /i/-, /e/-, /a/-, /o/-, /u/-activation. The dashed and small solid outlined boxes indicate phoneme regions, i.e. an activation of one phonemic state above 80% (dashed = /i/; solid = /e/; dotted = /a/; dash-dotted = /o/; dash-dott-dotted = /u/). The horizontal grey lines indicate the bark-scaled value of the first three formants representing the vocalic auditory state (bottom = 3 bark; top = 15 bark). The dark grey dot indicates the motor and proprioceptive representation of the vocalic tongue position (vertical = front – back; horizontal = low – high). The bold outlined boxes represent neurons which were activated by the vocalic stimulus continuum ("stimulus neurons"). The numbers within these boxes equals the stimulus number within the vocalic stimulus continuum.

need to be at least in a "next but one" relation, i.e. a maximum distance of one intermediate (non-stimulus) neuron is allowed between two neighboring neurons within a cluster. If the distance is greater than that, these two neurons are not members of the same cluster. (ii) A cluster needs to comprise at least 3 neurons.

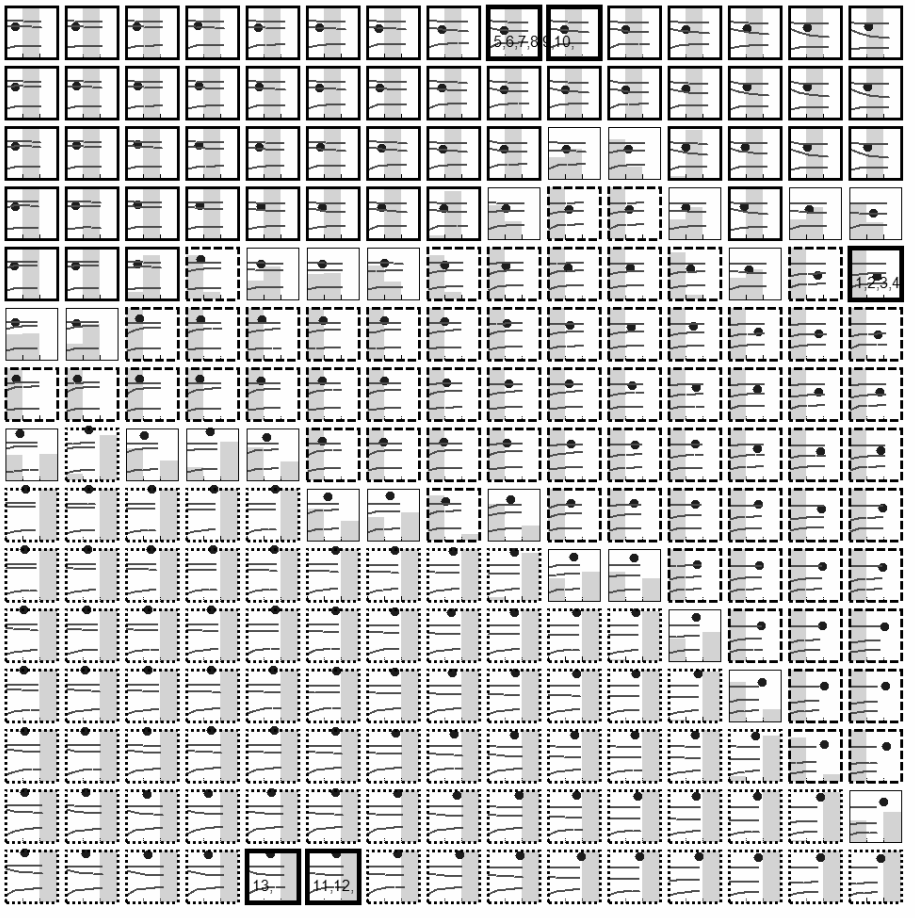

Fig. 3. Phonemic, motor, and sensory states represented by each neuron of the phonetic map for CV-syllables (model instance 11). Each box represents a neuron of the phonetic map for CV-syllables. Light grey bars represent the degree of activation of a phonemic state (from left to right: degree of /b/-, /d/-, /g/-activation. The dashed and small solid outlined boxes indicate phoneme regions, i.e. an activation of one phonemic state above 80% (dashed = /b/; solid = /d/; dotted = /g/). The horizontal grey trajectories indicate the bark-scaled value of the first three formants representing the auditory state of the syllable (bottom = 3 bark; top = 15 bark). The dark grey dot indicates the motor and proprioceptive representation of the final vocalic tongue position within the CV syllable (vertical = front – back; horizontal = low – high). Motor states can only be produced by those neurons which reach a phonemic activation above 80%. In these cases the bar represents the primary consonantal articulator (i.e. labial, apical or dorsal). The bold outlined boxes represent neurons which were activated by the consonantal stimulus continuum ("stimulus neurons"). The numbers within these boxes represent the stimulus number within the consonantal stimulus continuum.

The cluster analysis of all 20 virtual listeners indicates that for 19 virtual listeners, the V-part of the phonetic map exhibits one stimulus cluster covering all three phoneme regions (i.e. the /i/-, /e/-, and /a/-region). This can be interpreted as a continuous distribution of stimulus neurons over the three vocalic phoneme regions. Only one instance exhibits a stimulus cluster, which does not cover all three vocalic phoneme regions (Tab. 1). Furthermore for 11 (of 20) virtual listeners, the CV-part of the phonetic map clearly displays three clusters, each associated with a single phoneme region (i.e. /b/-, /d/-, and /g/-region). This can be interpreted as a clear case for clustering of stimulus neurons with respect to phoneme regions. For the remaining 9 virtual listeners, the CV-part of the phonetic map includes clusters covering more than one phoneme region (Tab. 1). Moreover in the CV-part of the phonetic map, 33 stimulus clusters were identified in total over all 20 virtual listeners, which can be associated with a single phoneme region. In contrast in the case of the V-part of the phonetic maps this only occurs for one stimulus cluster in total over all 20 virtual listeners (see Tab. 1).

Table 1. Results of a cluster analysis for the CV- and V-parts of the phonetic maps of all 20 virtual listeners (instances). The amount of instances, which exhibit the expected clustering, are indicated by bold letters. (CL = cluster; PR = phoneme region)

Type of instance	**amount of instances**		**amount of CL's associated with a single PR**	
	/CV/	**/V/**	**/CV/**	**/V/**
1 CL covering 3 PR	1	**19**	0	0
1 CL covering 2 PR	8	0	0	0
1 CL covering 1 PR	**11**	1	33	1
Total	20	20	**33**	**1**

5 Discussion

Our results indicate a stronger tendency towards a clustering of stimulus neurons in the case of the CV-part of phonetic maps than in the V-part. This result underlines that CV-stimuli (/C/ = /b/, /d/, or /g/) are perceived categorically while V-stimuli are perceived less categorically (more continuously) and this result is in accordance with the results of behavioral identification and discrimination experiments done by these virtual listeners [10]. It would be important now to create brain imaging experiments which support (or contradict) these results. But cortical phoneme regions seem to be very small, which makes these experiments very difficult [14].

In accordance with other approaches for modeling categorical perception [15, 9] our approach stresses the importance of *learning* and *neural self-organization* in order to reach typical features of categorical perception. But beyond other approaches our model stresses the importance of a supramodal "phonetic" level of self-organization which – beside linguistic information – takes into account sensory *and* motor information in parallel. We will demonstrate in further experiments that in the case of perception for place of articulation (labial – apical – dorsal) categories may emerge directly from anatomical facts and thus, infants just need (self-)babbling training in order to reach categorical perception of place of articulation. Since articulators (i.e. our hardware)

developed during evolution our model also delivers arguments for an evolution of categorical perception [16] at least for the perception of place of articulation.

Acknowledgments. This work was supported in part by the German Research Council DFG grant Kr 1439/13-1 and grant Kr 1439/15-1.

References

[1] Liberman, A.M., Harris, K.S., Hoffman, H.S., Griffith, B.C.: The discrimination of speech sounds within and across phoneme boundaries. Journal of Experimental Psychology 54, 358–368 (1957)
[2] Fry, D.B., Abramson, A.S., Eimas, P.D., Liberman, A.M.: The identification and discrimination of synthetic vowels. Language and Speech 5, 171–189 (1962)
[3] Eimas, P.D.: The relation between identification and discrimination along speech and non-speech continua. Language and Speech 6, 206–217 (1963)
[4] Mattingly, I.G., Liberman, A.M., Syrdal, A.K., Halves, T.: Discrimination in speech and nonspeech modes. Cognitive Psychology 2, 131–157 (1971)
[5] Burns, E.M., Campbell, S.L.: Frequency and frequency-ration resolution by possessors of absolute and relative pitch: Examplex of categorical perception? Journal of the Acoustical Society of America 96, 2704–2719 (1994)
[6] Mirman, D., Holt, L.L., McClelland, J.L.: Categorization and discrimination of non-speech sounds: differences between steady-state and rapidly-changing acoustic cues. Journal of the Acoustical Society of America 116, 1198–1207 (2004)
[7] Poeppel, D., Guillemin, A., Thompson, J., Fritz, J., Bavelier, D., Braun, A.R.: Auditory lexical decision, categorical perception, and FM direction discrimination differentially engage left and right auditory cortex. Neuropsychologia 42, 183–200 (2004)
[8] Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T., Medler, D.A.: Neural substrates of phonemic perception, vol. 15, pp. 1621–1631 (2005)
[9] Beer, R.D.: The dynamics of active categorical perception in an evolved model agent. Adaptive Behavior 11, 209–243 (2003)
[10] Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C.: Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793–809 (2009)
[11] 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.) COST Action 2102. LNCS (LNAI), vol. 4775, pp. 174–189. Springer, Heidelberg (2007)
[12] 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)
[13] Kohonen, T.: Self-Organizing Maps. Springer, Berlin (2001)
[14] Obleser, J., Boecker, H., Drzezga, A., Haslinger, B., Hennenlotter, A., Roettinger, M., Eulitz, C., Rauschecker, J.P.: Vowel sound extraction in anterior superior temporal cortex. Human Brain Mapping 27, 562–571 (2006)
[15] Damper, R.I., Harnad, S.R.: Neural network models of categorical perception. Perception and Psychophysics 62, 843–867 (2000)
[16] Kuhl, P.K.: Early language acquisition: cracking the speech code. Nature Reviews Neuroscience 5, 831–843 (2004)