Towards the Acquisition of a Sensorimotor Vocal Tract Action Repository within a Neural Model of Speech Processing

Bernd J. Kröger[1], Peter Birkholz[1], Jim Kannampuzha[1],
Emily Kaufmann[2], and Christiane Neuschaefer-Rube[1]

[1] Department of Phoniatrics, Pedaudiology, and Communication Disorders,
University Hospital Aachen and RWTH Aachen University, Aachen, Germany
{bkroeger,pbirkholz,jkannampuzha,cneuschaefer}@ukaachen.de
[2] Human Technology Centre, RWTH Aachen University, Aachen, Germany
kaufmann.emily@gmail.com

Abstract. While a mental lexicon stores phonological, grammatical and semantic features of words, a vocal tract action repository is assumed to store inner motor and sensory representations of speech items (i.e. the sounds, syllables and words) of the speaker's native language. On the basis of a neural model of speech processing, which comprises important cognitive and sensorimotor aspects of speech production, perception, and acquisition (Speech Commun 51, 793–809, 2009), this paper will outline how a sensorimotor vocal tract action repository can be acquired in a self-organizing neural network structure which is trained using unsupervised associative learning.

Keywords: Speech actions, neural model, speech production, speech perception, speech acquisition, mental lexicon, neural network, self-organization.

1 Introduction

Neural models of speech processing aim to account for cognitive, sensory, and motor aspects of speech production and perception ([1], [2], and [3]). While the *mental lexicon* plays a major role as a repository for the cognitive linguistic description of words [4], a *mental syllabary* is presumed to be the central repository for the sensory and motor representation of frequent syllables ([4], [5], and [6]). A comparable module, which we will call the *sensorimotor vocal tract action repository*, is presumed to represent the mentally syllabary in our approach [3]. The central structural feature of this module is a self-organizing map (a *phonetic map* or hypermodal action map: P-MAP). This map associates the motor, sensory, and phonemic states of the most frequent syllables. Our model has already been tested for a limited *model language* data set comprising a simple vowel and consonant system with 45 CV- and 20 CCV-syllables (V = vowel, C = consonant; [7] and [8]). In this paper, a simulation experiment will be described in which the system acquired a basic set of 200 syllables of a *natural language*, i.e. Standard German in the case of this study.

A. Esposito et al. (Eds.): Communication and Enactment 2010, LNCS 6800, pp. 287–293, 2011.

2 The Neural Model

Our neural model comprises two knowledge repositories, i.e. the mental lexicon and the action repository, as well as modules for neuromuscular and perceptual processing (Fig. 1). Phonological and semantic processing modules outside the mental lexicon are not yet integrated into the model. Word production starts with local neural activations within the *semantic self-organizing map* (S-MAP). Here, one model neuron represents one lexical item, i.e. one word. This neural activation leads to a co-activation of a distributed neural activation pattern, representing the semantic state of that word. The S-MAP is also connected with the *phonetic self-organizing map* (P-MAP), leading to a co-activation of those model neurons within the P-MAP which represent the syllables of that word. Thus, phonemic states, motor plans, and internal sensory states are also co-activated for these syllables ([3] and [9]). This activation triggers the execution (i.e. articulation) of the word. Then, the still-activated inner sensory states of each syllable can be compared with their external sensory states using the articulatory-acoustic model (*sensorimotor feedback loop*). A detailed babbling and imitation training which establishes the phonetic map and the neural associations with the motor plan and sensory maps has been described for V- and CV-syllable states [3] and for V-, CV- and CCV-syllable states; see [7] and [8].

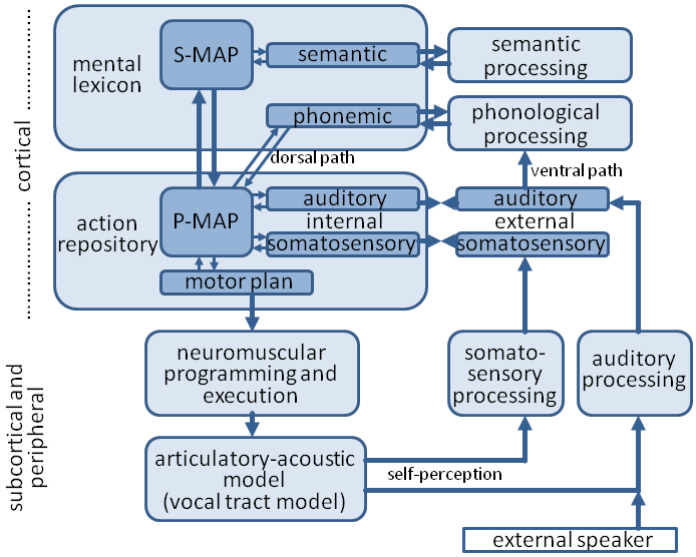

Fig. 1. Structure of the neural model of speech processing. Light blue boxes indicate processing modules; dark blue boxes indicate self-organizing maps (S-MAP and P-MAP) or neural state maps, i.e. the semantic, phonemic, auditory, somatosensory, and motor plan state map.

3 Method: Training the Model

A word and syllable list was assembled based on our corpus of Standard German children's books, which comprises transcriptions of 40 books targeted to children between one and six years of age. This corpus comprises 6513 sentences and 70512 words in total, with morphologically distinct forms of the same word counted as separate words (e.g. 'kleine' and 'kleinen', two forms of the word 'klein', meaning *small*, which are used with nouns of different grammatical genders and in different grammatical cases). A further analysis revealed that the corpus comprises 8217 different words, which is assumed to approximately represent a 6-years-old child's mental lexicon (Tab. 1). These words were phonetically transcribed using phonetic transcription rules for Standard German [10]. There were 4763 different syllables found in the transcription, of which 2139 syllables can be defined as *frequent syllables*: 96% of the corpus sentences can be produced using only these 2139 syllables (Tab. 2).

The *200 most frequent syllables*, including phonetic simplifications which typically occur in children's word production (e.g. elisions and assimilations of sounds [12]), comprise CV-, CVC-, CVCC-, CCV-, and CCVC-syllables. Typical frequent CV-syllables comprise the consonants [t, ʔ, g, n, d, z, b, l, r, s, h, m, k, l, f, v, ç, ʃ, p, x, j] together with the vowels [ə, ʊ, ɪ, aɪ, a, iː, ɛɐ, aʊ, eː, ɛ, ɐ, ɔ, uː, aː, eː, oː, aɐ, oɐ, iɐ, yː]. Typical frequent CVC-syllables are [ˈʔʊn, ˈdas, ˈʔaɪn, ˈʔɪs, ˈʔɛs, ˈhat, ˈʔaʊf, ˈmɪt, ˈʔan, ˈnɪç, ˈzɪç, ˈzaːt, ˈʔɪç, ˈʔɪm, ˈʔaʊs] ([ˈ] indicates a stressed syllable). Typical frequent CVCC-syllables are [ˈʔʊnt, ˈʔɪst, ˈʔals, ˈnɪçt, ˈzaːkt, ˈmoːnt, ˈʔalt, ˈkɔmt]. Typical frequent CCV-syllables are [ˈtsuː, tsə, ˈklaɪ, ˈtsaɪ, ˈʃpiː]. Typical frequent CCVC-syllables are [ˈtsən] and [ˈʃtɛɐn].

Table 1. The ten most frequent words in the categories noun, verb, adjective/adverb and other (i.e. pronouns and particles; particles comprise prepositions, conjunctions, and interjections [11]), in our corpus of Standard German; N = frequency of occurrence of that word.

Nouns	**N**	**Verbs**	**N**	**Adj./Adv.**	**N**	**Others**	**N**
"Mama" (mom)	392	"ist" (is)	793	"kleine" (little)	287	"und" (and)	2367
"Bär" (bear)	278	"hat" (has)	448	"mehr" (more)	126	"die" (the)	1678
"Papa" (dad)	235	"sagt" (says)	413	"schnell" (fast)	90	"der" (the)	1644
"Mond" (moon)	217	"war" (was)	246	"viel" (much)	75	"sie" (she/it)	1391
"Kinder" (children)	190	"kann" (can)	184	"kleinen" (little)	74	"das" (the)	891
"Katze" (cat)	147	"wird" (will be)	159	"fest" (fixed)	67	"den" (the)	831
"Frau" (wife)	145	"will" (want)	156	"genau" (exactly)	60	"ein" (a)	781
"Bett" (bed)	106	"sagte" (said)	131	"großen" (large)	59	"er" (he)	777
"Mädchen" (girl)	105	"muss" (must)	120	"einfach" (simple)	58	"es" (it)	764
"Wasser" (water)	104	"sieht" (sees)	112	"große" (large)	58	"in" (in)	616

Table 2. Number N of most frequent syllables occurring at least M times within the corpus and percentage of text or speech which can be produced using only these syllables

Number N of most frequent syllables	Minimum number M of instances of each of these N most frequent syllables	Percentage of sentences within the corpus which can be produced using the N most frequent syllables
477	>= 40	75%
856	>= 20	85%
1396	>= 10	91%
2139	>= 5	96%
2843	>= 3	98%
3475	>= 2	99%
4763	>= 1	100%

The training of the phonetic map (P-MAP) was done in two steps. First, the training set (comprising phonemic, auditory, and motor plan states) was established for the 200 most frequent syllables. This was done by (i) choosing one acoustic realization of each syllable produced by one speaker of Standard German (33 years old, male), who uttered a selection of the sentences listed in the children's book corpus, and (ii) applying an articulatory-acoustic re-synthesis method [13] in order to generate the appropriate motor plans. Each *auditory state* is based on the acoustic realization and is represented in our model as a *short-term memory spectrogram* comprising 24 × 30 neurons, where 24 rows of neurons represent the 24 critical bands (20 to 16000 Hz) and where 65 columns represent successive time intervals of 12.5 ms each (overall length of short-term time interval: 812.5 ms). The degree of activation of each neuron represents the spectral energy within a time-frequency interval. Each *motor plan state* is based on the motor plan generated by our re-synthesis method [13] and is represented in the neural model by a vocal tract action score as introduced in [14]. The score is determined by considering (i) a specification of the temporal organization of vocal tract actions within each syllable (i.e. 11 action rows over the whole short-term time interval: 11 × 65 neurons) and (ii) a specification of each type of action (4 × 17 for consonantal and 2 × 15 for vocalic actions; assuming CCVCC as the maximally complex syllable structure). Each *phonemic state* is based on the discrete description of all segments (allophones) of each syllable: 159 neurons in total.

In the second step, this *syllabic sensorimotor training set*, covering the 200 most frequent syllables, was applied in order to train three P-MAPS of different sizes i.e. self-organizing neuron maps with 15 × 15, 20 × 20, and 25 × 25 neurons, respectively. 5000 incremental training cycles were computed using standard training conditions for self-organizing maps [3]. The training of the P-MAP can be called *associative training* since phonemic, motor, and sensory states are presented synchronously to the network for each syllable. Each cycle comprised 703 incremental training steps, and each syllable was represented within the training set proportionally to the frequency of its occurrence in the children's book corpus; i.e. the most frequent syllable occurred 25 times per training cycle, while the least frequent syllable (number 200 in the ranking) occurred one time per cycle. Thus, the least-frequent syllable appeared 5000 times in total, and the most frequent syllable appeared 125000 times in total in the training.

4 Results

Our simulation experiments indicate that a P-MAP comprising at least 25 × 25 neurons is needed in order to represent all 200 syllables. 158 syllables were represented in the 15 × 15 phonetic map, and 176 syllables were represented in the 20 × 20 map (see Fig. 2) after training was complete.

'dEA	'dE	'mI	'mIt		'nIt		'?aUf	'?a	'?al		'?IC	'?IC	'?I	'?I	'?I	'?Is	'?Is		'?Unt
'dEA		'mI	'mI		'nI	'nIC		'?am	'?a	'?a:			'?Im	'?I	'?It		'?Is	'?Ist	'?U
			'bI	'bIs		'nIC	'nICt	'?a	'?a	'?a	'?aI	'?aI		'?In		'?Un	'?Un		'?U
'bE		'SE	'SI		lIC	lI		'?aUs	'?a	'?as		'?aI		'?In		'?Un		'?Um	'?U
	'bEA			'kI		lI	lI				'zIC		'?aIn		'?an		'tsUm		'?Ut
fEA		'de:n	'de:n	'kI	'kIn		'?aU	'baU		'zInt	'zIC	n@n	'?aIn	ts@n				'mUs	
'fyA					'hIn		'?aU		'zI		n@n	n@n		t@n	t@n		pa:		ma:
'foA		d@	d@	s@n		C@		'?aUx		n@m	'di:		ni:	ni:		'pa			ma
	'de:		d@n		h@		C@n		n@		'di:	'di:		ni:	'ma:l	'ma	'ma	ma	ma
'kO				f@n		S@		l@n	n@	b@n	'di:		pi:		'ma:l	'ma:l		ta	
'kO	'hOI			k@n	f@	p@	l@	l@	b@	b@n		'du:		li:	'li:		t@	t@	t@t
'kOn				k@		z@			b@	b@					ts@	ts@	t@	t@l	
	'fO		nOk		raU			laI		vaI		'taI		'tu:			'va		'das
m@		dA		rI	ra	ra:	ra:		'kaI		'baI		'tsu:		'vaA		'va	'vas	
	mA		r@		raI			'klaI		'zaI		'vi:		'nuA		'ga	'gan		'ka
nA	lA	g@l		r@n		'?Es			'zi:		'viA		'vI		'la		'kan		'kat
	gA	g@	g@				'?i:	'?i:	'zi:	'zi:		'fi:		'ha	'Sa	'ja		'dan	'da
bA		g@n	g@	vEn		'?E		'?o:		'Si:	'Si:				'ha:		'hat		'da
			tE		'?En	'?Et	'?y:		'So:n			'ri:	'gro:	'na:	'za:	'za:k			'da
tA	tA	tI		'?EA	'?EA	'?iA		'mo:nt		jo:	ro:	'vo:	'zo:		'za:t	'za:	'za:kt		'da:

Fig. 2. Organization of the 20 × 20 neuron P-MAP. Each box represents a neuron within the self-organizing neural map. A syllable appears only if the activation of its phonemic state is greater than 80% of maximum activation.

While most of the syllables are represented by only one neuron in the 15 × 15 map, approximately the 100 most frequent syllables are represented by two or more neurons in the 20 × 20 and 25 × 25 maps. This allows the map to represent more than one realization for each of these syllables (e.g. [ˈda] is represented by 3 neurons, while [ˈdan] and [ˈkan] are represented by only one neuron each in the 20 × 20 map:

see Fig. 2). It should be noted that the syllables in Figure 2 are loosely ordered with respect to syllable structure (e.g. CV vs. CCV or CVC), vowel type (e.g. [i] vs. [a]) and consonant type (e.g. plosive vs. fricative or nasal).

5 Discussion

Our neural model of speech processing as developed thus far is capable of simulating the basic processes of acquiring the motor plan and sensory states of frequent syllables of a natural language by using unsupervised associative learning. This process is illustrated here on the basis of our Standard German children's book corpus that 96% of fluent speech can be produced using only the 2000 most frequent syllables. These frequent syllables are assumed to be produced directly by activating stored motor plans, without using complex motor processing routines.

In our neural network model, the sensory and motor information about frequent syllables is stored by the dynamic link weights of the neural associations occurring between a self-organizing P-MAP and neural state maps for motor plan, auditory, somatosensory, and phonemic states. Thus, a neuron within the P-MAP represents a syllable, which – if activated – leads to a syllable-specific activation pattern within each neural state map. These neural activations represent "internal speech" or "verbal imagery" [15], i.e. "how to articulate a syllable" (motor plan state), "what a syllable sounds like" (auditory state), and "what a syllable articulation feels like" (somatosensory state), without actually articulating that syllable.

While in earlier experiments our simulations were based on an artificial and completely symmetric *model language*, comprising five vowels [i, e, ɛ, o, a] and nine consonants [b, d, g, p, t, k, m, n, l] and all combinations of vowels and consonants as CV-syllables and all combinations of four CC-clusters [bl, gl, pl, kl] with all vowels as CCV-syllables, this paper gives the first results of simulation experiments based on a *natural language*, i.e. based on the 200 most frequent syllables of Standard German as they occur in our children's book corpus, including phonetic simplifications which typically occur in children's word production. While syllables are strictly ordered with respect to phonetic features in the P-MAP in the case of the model language (see [3], [7], and [8]), we can see here that syllables are ordered more "loosely" in the case of a natural language. This is due to the fact that natural languages are less symmetrical than the model language due to the gaps in syllable structure which are present in a natural language, i.e. not all combinations of vowels and consonants are equally likely to occur in a natural language as they are in a model language.

Furthermore, our simulations indicate that the representation of 200 syllables within the P-MAP requires a minimum map size of 25 × 25 neurons. Phonetic maps of 15 × 15 or 20 × 20 neurons were not capable of representing all 200 syllables. In order to be able to account for complete acquisition of a language, more than 200 syllables (up to 2000) must be included in the training set, so the size of the P-MAP and the S-MAP must be increased before this will be possible (cf. [9]).

Acknowledgments. We thank Cornelia Eckers and Cigdem Capaat for building the corpus. This work was supported in part by the German Research Council (DFG) grant Kr 1439/13-1 and grant Kr 1439/15-1 and in part by COST-action 2102.

References

1. 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)
2. Guenther, F.H., Vladusich, T.: A neural theory of speech acquisition and production. Journal of Neurolinguistics (in press)
3. Kröger, B.J., Kannampuzha, J., Neuschaefer-Rube, C.: Towards a neurocomputational model of speech production and perception. Speech Communication 51, 793–809 (2009)
4. Levelt, W.J.M., Roelofs, A., Meyer, A.: A theory of lexical access in speech production. Behavioral and Brain Sciences 22, 1–75 (1999)
5. Levelt, W.J.M., Wheeldon, L.: Do speakers have access to a mental syllabary? Cognition 50, 239–269 (1994)
6. Wade, T., Dogil, G., Schütze, H., Walsh, M., Möbius, B.: Syllable frequency effects in a context-sensitive segment production model. Journal of Phonetics 38, 227–239 (2010)
7. Kröger, B.J.: Computersimulation sprechapraktischer Symptome aufgrund funktioneller Defekte. Sprache-Stimme-Gehör 34, 139–145 (2010)
8. 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. Plural Publishing, San Diego (in press)
9. Li, P., Farkas, I., MacWhinney, B.: Early lexical development in a self-organizing neural network. Neural Networks 17, 1345–1362 (2004)
10. Kohler, W.: Einführung in die Phonetik des Deutschen. Erich Schmidt Verlag, Berlin (1995)
11. Glinz, H.: Deutsche Syntax. Metzler Verlag, Stuttgart (1970)
12. Ferguson, C.A., Farwell, C.B.: Words and sounds in early language acquisition. Language 51, 419–439 (1975)
13. Bauer, D., Kannampuzha, J., Kröger, B.J.: Articulatory Speech Re-Synthesis: Profiting from natural acoustic speech data. In: Esposito, A., Vích, R. (eds.) Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions. LNCS (LNAI), vol. 5641, pp. 344–355. Springer, Heidelberg (2009)
14. Kröger, B.J., Birkholz, P., Lowit, A.: Phonemic, sensory, and motor representations in an action-based neurocomputational model of speech production (ACT). In: Maassen, B., van Lieshout, P. (eds.) Speech Motor Control: New Developments in Basic and Applied Research, pp. 23–36. Oxford University Press, Oxford (2010)
15. Ackermann, H., Mathiak, K., Ivry, R.B.: Temporal organization of "internal speech" as a basis for cerebellar modulation of cognitive functions. Behavioral and Cognitive Neuroscience Reviews 3, 14–22 (2004)