TRAINING EINER SICH SELBST ORGANISIERENDEN KARTE IM NEUROBIOLOGISCHEN SPRACHVERARBEITUNGSMODELL MSYL

Jim Kannampuzha, Cornelia Eckers und Bernd J. Kröger

Klinik für Phoniatrie, Pädaudiologie und Kommunikationsstörungen
Universitätsklinikum Aachen und RWTH Aachen University
{jkannampuzha, ceckers, bkroeger}@ukaachen.de

Kurzfassung: Qualitative Sprachproduktions und -wahrnehmungsmodelle gehen von der Existenz eines mentalen Silbenspeichers aus. Dieser assoziert und ordnet die motorischen, auditiven und phonemischen Repräsentationen von häufig produzierten Silben miteinander. In dieser Arbeit wird eine quantitative Modellierung des Lernvorgangs zum Aufbau eines mentalen Silbenspeichers vorgestellt. Dieses Modell (MSYL) lernt die Assoziationen zwischen den obengenannten Repräsentationen eines Items in einer Selbstorganiserenden Karte (SOM). Das dem Lernen zugrunde liegende Silbenkorpus besteht aus den 200 häufigsten Silben, die aus einem Kinderbuch-Satzkorpus gewonnen wurden. Nach dem Training erhält man eine SOM, an der sich die Eigenschaft der kortikalen Plastizität zeigen lässt. Man sieht wie mit fortschreitendem Training die Anzahl der von der SOM assoziierten Items steigt und eine Ordnung einnimmt.

1 Motivation

Qualitative Sprachproduktionsmodelle [7] gehen von der Existenz eines mentalen Silbenspeichers aus, der Informationen zu häufig gesprochenen Silben enthält. Aktuelle quantitative Sprachproduktionsmodelle sind zum einen das DIVA Modell [2] und zum anderen das neurophonetische Sprachproduktions und -wahrnehmungsmodell von Kröger [6]. Während das Erstere den Schwerpunkt auf die Modellierung der Rückkopplungsmechanismen legt, konzentriert sich das neurophonetische Modell auf den Aufbau und die Organisation des mentalen Silbenspeichers. In dieser Arbeit wird die quantitative Modellierung des Lernvorgangs zum Aufbau eines mentalen Silbenspeichers mit Hilfe einer Selbstorganisierden Karte (SOM) beschrieben. Die Struktur dieses Modells MSYL (Abb. 1) wurde auf der Basis neurophysiologischer Literatur [6] erstellt. Dabei sind alle Neuronen der SOM mit allen Neuronen der Zustandskarten verbunden. Die SOM assoziert die phonemischen, auditiven, somatosensorischen und motorischen Zustände miteinander. Die SOM mit ihren Verbindungsgewichten stellt das Langzeitgedächtnis dar. Die Zustandskarten hingegen bilden nur einen Zeitraum von bis zu 800 ms ab und repräsentieren das Kurzzeitgedächtnis. Die somatosensorische Karte wurde noch nicht implementiert. Das Modell hat drei unterschiedliche Arbeitsmodi (Perzeptions-, Produktions- und Lernmodus). Die durchgezogenen Pfeile zeigen die Aktivierungsrichtung im Lernmodus. Die gestrichelten Pfeile die Aktivierungsrichtung bei der Perzeption. Dabei ist der Anfangspunkt entweder ein externer Sprecher oder das Vokaltraktmodell, was die Eigenwahrnehmung des selbst Gesprochenen modelliert, und der Endpunkt die phonemische Karte, hierbei werden die somatosensorischen und motorischen Karten koaktiviert (Perzeptionsmodus). Die gepunkteten Pfeile zeigen die Aktivierungsrichtung bei der Produktion. Der Anfangspunkt ist hier die phonemische Karte und der

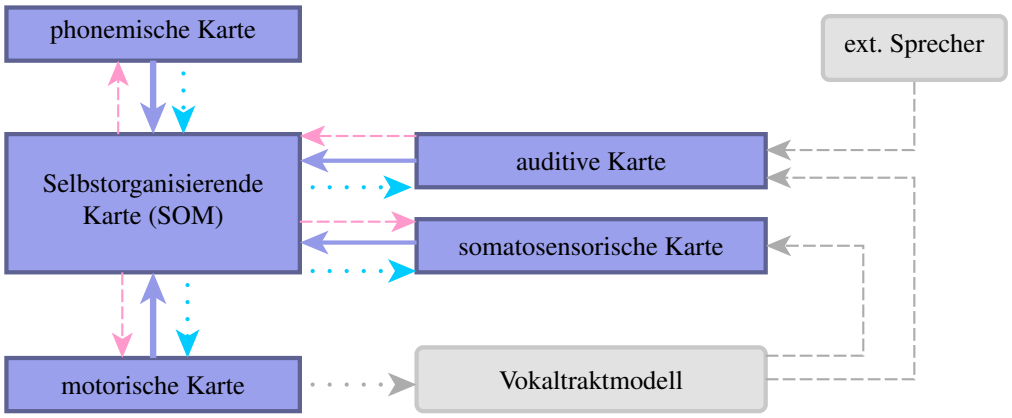

Abbildung 1 - Die Struktur des Modells MSYL.

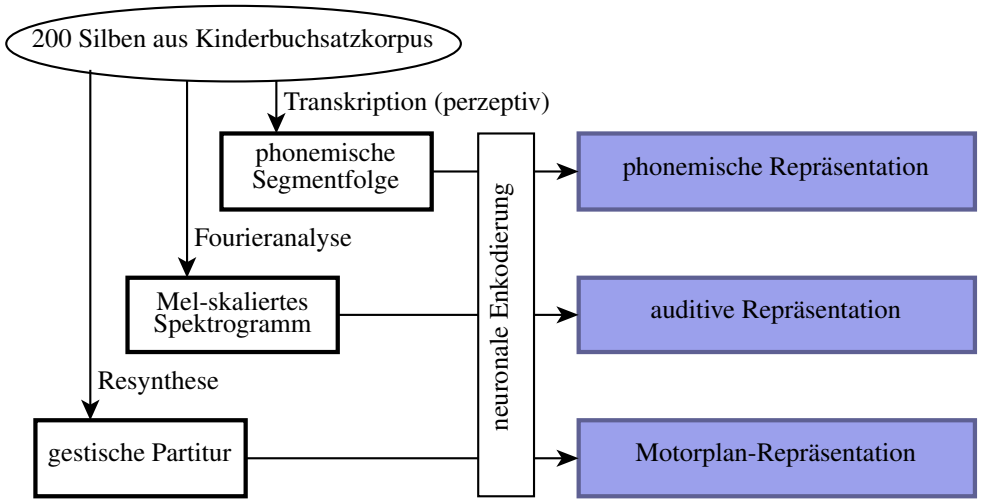

Abbildung 2 - Die Umwandlung eines Elements aus dem Silbenkorpus in ein Trainingsitem der Trainingsmenge

Endpunkt das Vokaltraktmodell (Produktionsmodus). Dabei werden auch die auditive und die somatosensorische Karte zur Kontrolle koaktiviert.

2 Methode

2.1 Erstellung des Silbenkorpus

Das zum Training des Modells MSYL verwendete Silbenkorpus umfasst die 200 häufigsten Silben auf der Basis eines Kinderbuch-Satzkorpus, welches 40 Bücher mit insgesamt 6513 Sätzen und 70512 Wörtern inklusive Wiederholungen umfasst [5]. Diese Wörter wurden von einem 32-jährigen männlichen Sprecher im Satzzusammenhang eingesprochen und anschließend auf Wortebene geschnitten und auf Silbenebene mit der zugehörigen phonemischen Segmentfolge in SAMPA-Notation [8] annotiert. Für jede Silbe hat man aus einer Realisierung eines Wortes, das diese Silbe enthält, durch Fourieranalyse das Mel-skalierte Amplitudenspektrogramm und durch Resynthese [1] die gestische Partitur als spezifische Form des Motorplans der Silbe erstellt (Abb. 2 links). Das Silbenkorpus besteht also aus 200 Items und jedes Item besteht aus einer phonemischen Segmentfolge, einem Spektrogramm und einer gestischen Partitur. Die neuronale Enkodierung der Items wird im folgenden Abschnitt näher beschrieben.

Onset	m	b	p	n	d	t	N	g	k	f	v	s	z	S	Z	C	x	h	l	r	j	?	w
KS1 j	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
KS2 -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KS3 -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Nukleus	i	y	I	Y	e	2	E	9	a	o	O	u	U	6	@	:
VS1 E	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
VS2 -	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Offset	m	b	p	n	d	t	N	g	k	f	v	s	z	S	Z	C	x	h	l	r	j	?	w
KS4 t	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
KS5 s	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
KS6 t	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

BET ’	1

Tabelle 1 - Die neuronale Darstellung des phonemischen Zustands der Silbe ['jEtst] (“jetzt”). KS und VS stehen für konsonantisches bzw. vokalisches Segment.

2.2 Neuronale Enkodierung eines Items

Für das Training im SOM müssen die Items neuronal enkodiert werden. In folgenden Unterabschnitten werden die Kodierungen der phonemischen Segmentfolge, des Spektrogramms und des Motorplans beschrieben (Abb. 2 rechts). Diese neuronalen Repräsentationen bilden die Zustände, die die phonemische, auditive und motorische Karte im Modell annehmen kann. Die Trainingsmenge besteht dann aus den neuronalen Repräsentationen aller Items aus dem Silbenkorpus. Die Anzahl der Wiederholungen wurden anschließend dergestalt linear skaliert, dass das Item mit der geringsten Häufigkeit im Korpus nur einmal vorhanden war, daraus ergeben 26 Wiederholungen für das häufigste Item.

2.2.1 Kodierung der phonemischen Segmentfolge

In dieser Studie nehmen wir für eine Silbe eine Maximalanzahl von 3 konsonantischen Segmenten im Silbenonset (KS1, KS2, KS3) und Silbenoffset (KS4, KS5, KS6) an und maximal 2 vokalische Segmente – für die Darstellung aller Vokale und Diphthonge – im Silbennukleus (VS1, VS2). Jede der 3 Positionen im Onset und Offset kann von 23 konsonantischen Segmenten und jede der 2 Positionen im Nukleus von 16 vokalischen Segmenten eingenommen werden (Tab. 1). Die neuronale Repräsentation eines konsonantischen bzw. eines vokalischen Segments besteht somit aus 23 bzw. 16 Neuronen, wobei jedes Neuron ein Segment darstellt. Ein Neuron wird auf den Wert 1 gesetzt (maximale Aktivierung), wenn das zugehörige Segment in der Segmentfolge an entsprechender Position vorhanden ist, alle anderen Neuronen werden auf den Wert 0 (keine Aktivierung) gesetzt. Ist kein Segment vorhanden werden alle Neuronen auf 0 gesetzt. Weiterhin wird ein Neuron für die Kodierung der Betonung angesetzt. Die vollständige neuronale Repräsentation der Segmentfolge besteht dann aus 171 Neuronen.

2.2.2 Kodierung des Mel-skalierten Spektrogramms

Die neuronale Darstellung der auditiven Repräsentation eines Items entspricht einer diskretisierten Darstellung des Mel-skalierten Spektrogramms. Ein Zeitintervall von 800 ms wurde gewählt, da sich alle Silben in diesem Intervall unterbringen ließen. Für die Diskretisierung wurden Blöcke von 12,5 ms gewählt. Die Frequenzachse wurde entsprechend der Bark-Skala in 24 diskrete Bereiche unterteilt. Die Amplitudenwerte in einem Dynamikbereich von 120 dB wurden auf den Wertebereich zwischen 0 und 1 normiert (keine bis maximale Aktivie-

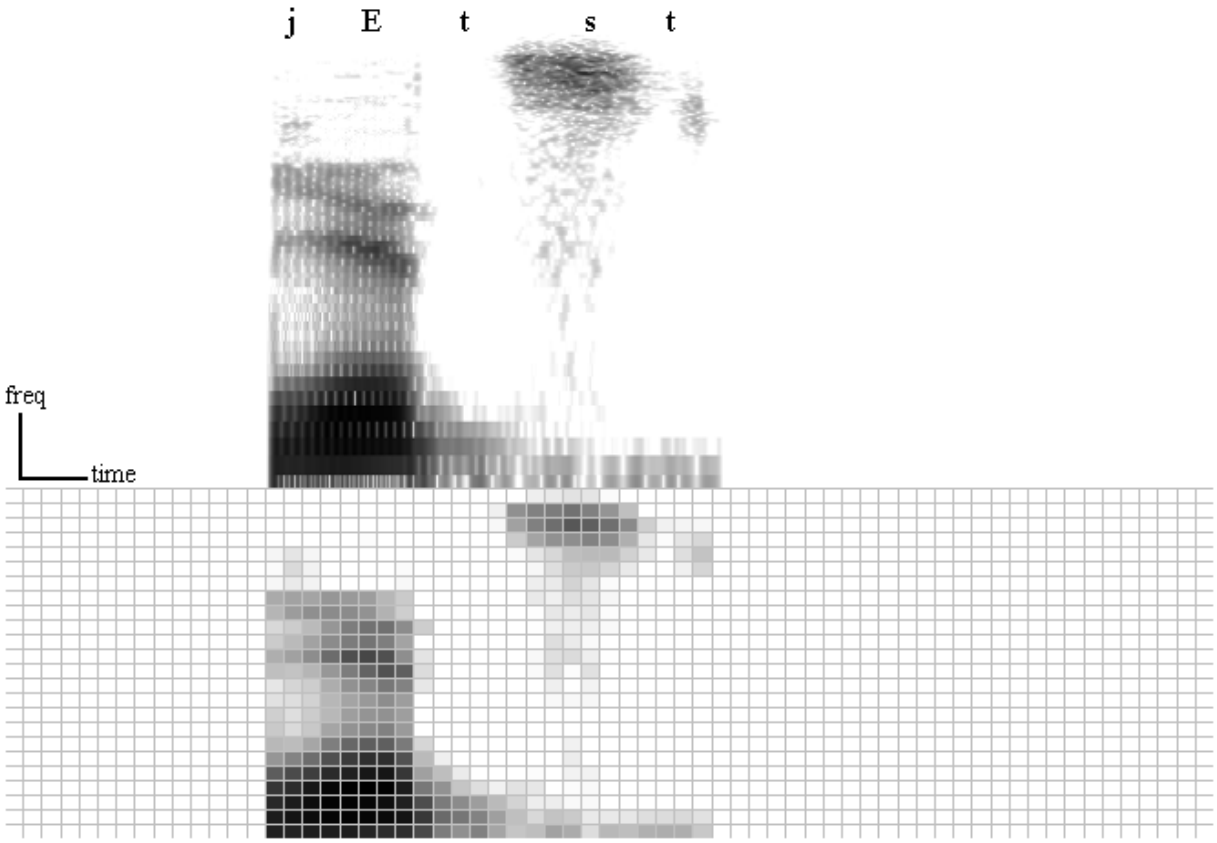

Abbildung 3 - Die auditive Repräsentation des Items ['jEtst] des Silbenkorpus (oben) und ihre neuronale Enkodierung wie in Abschnitt 2.2.2 beschrieben (unten). Die grauen Farben in der Darstellung der neuronalen Repräsentation des Spektrogramms stehen für den Aktivierungsgrad der Neuronen. Sie liegt zwischen 0 (keine Aktivierung, weiß) und 1 (maximale Aktivierung, schwarz).

rung). Man erhält dann ein 64×24 Felder großes Neuronengitter, das die auditive Zustandskarte darstellt. Abbildung 3 zeigt die neuronale Repräsentation des Spektrogramms für das Item ['jEtst] ("jetzt") des Silbenkorpus.

2.2.3 Kodierung der gestischen Partitur zum Motorplan

Der Motorplan liegt aufgrund der Resynthese zunächst in Form einer gestischen Partitur vor (Abb. 4 oben). Die gestische Partitur besteht aus einer Liste von Gesten nach [4]. Eine Geste wird beschrieben durch einen Gestentyp, ein Gestenlabel und einer Zeitfunktion des relativen Artikulator-Target-Abstandes, die die zeitliche Struktur der Geste darstellt (siehe gestrichelte Geste in der velopharyngalen Reihe). Sie gibt zu jedem Zeitpunkt den Grad der Annäherung an das Ziel (Target) der Geste, bezogen auf die Ausgangsposition, an. Der streng monoton steigende Teil ist die Annäherungsphase, darauf folgt die Targetphase und der streng monoton fallende Teil ist die Lösungsphase [5]. Der Gestentyp und das Gestenlabel kodieren die räumliche Struktur der Geste (Artikulator, Ort der Konstriktion). Es gibt 5 verschiedene Typen von Gesten, nämlich vokalische, konsonantische, velopharyngale, glottale und pulmonale Gesten. Diese werden in Abbildung 4 oben auf unterschiedlichen artikulatorischen Reihen ("tiers") angezeigt. Die Gestenlabel, mit welchen eine Geste markiert werden kann, hängen von dem Gestentyp ab. Die Label für konsonantische und vokalische Gesten sind in Tabelle 2 bzw. 3 aufgelistet. Velopharyngale Gesten können nur mit den Gestenlabeln 'open' und 'close' markiert

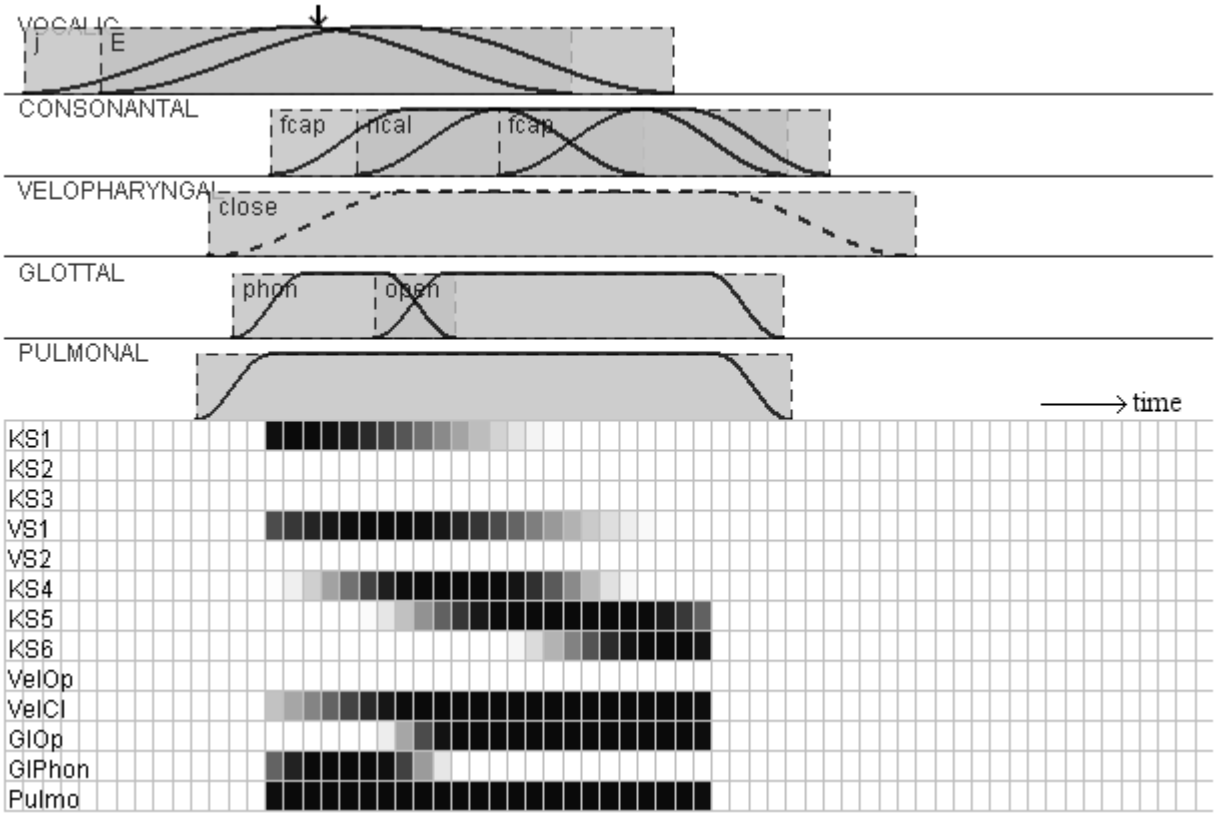

Abbildung 4 - Im oberen Teil sieht man die gestische Partitur einer Resynthese der Silbe ['jEtst]. Im unteren Teil ist die aus der gestischen Partitur generierte neuronale Repräsentation der zeitlichen Struktur zu sehen. Insgesamt ist ein Zeitbereich von 800 ms abgebildet. Der kleine Pfeil zeigt den Beginn des Silbenkerns an (siehe Abschnitt 2.2.4).

werden. Sie beschreiben eine offene bzw. geschlossene velopharyngale Enge. Glottale Gesten können zusätzlich mit dem Label 'phon' markiert werden. Sie werden für die Resynthese von stimmlosen Lauten (offene Stimmritze, 'open'), von dem glottalen Verschluss (fest verschlossene Strimmritze, 'close') und von stimmhaften Lauten (geschlossene Stimmritze mit Phonation, 'phon') benötigt. Pulmonale Gesten beschreiben den Lungendruck und haben kein Label. Die Gestenlabel werden in Abbildung 4 oben in der linken oberen Ecke jeder Geste angezeit.

Vor der neuronalen Enkodierung der gestischen Partitur wurden die Gesten je nach Gestentyp und je nach der Position in der Silbe (siehe Abschnitt 2.2.1) gruppiert. Dann ergeben sich die Gruppen KS1, KS2, KS3, VS1, VS2, KS4, KS5 und KS6, welche für die Positionen in der Silbe stehen. Gesten, die in der Resynthese die Segmente an diesen Positionen beschreiben, werden der jeweiligen Gruppe hinzugefügt. Weitere Gruppen sind VelOp und VelCl, die velopharyngale Öffnungs- und Verschlussgesten enthalten, GlOp und GlPhon, die glottale Öffnungs- und Phonationsgesten enthalten und die Gruppe Pulmo, die die Gesten für den Lungendruck enthält. Die glottale Verschlussgeste wird in eine der konsonantischen Gruppen zugeordnet. Für die neuronale Enkodierung der Zeitwerte wurde wieder ein Zeitraum von 800 ms betrachtet und in Blöcken von 12,5 ms eingeteilt, so dass 64 Neuronen diesen Zeitraum kodieren. Ein Neuron bildet dann den durschnittlichen Wert der Zeitfunktion des relativen Artikulator-Target-Abstandes in dem entsprechenden Zeitblock ab. Jede Gruppe wurde separat enkodiert, daraus ergeben sich 13×64 Neuronen für die zeitliche neuronale Enkodierung (Abb. 4 unten). Für die neuronale Enkodierung der räumlichen Struktur werden für die Gruppen KS1 bis KS6 zusätzlich Neuronen

Konsonantische Gestenlabel	bilabial	labio-dental	apikal/ alveolar	post-alveolar	palatal	dorsal/velar/ uvular	glottal
full closure (fc)	0 [fcla]	—	0 [fcap]	—	—	0 [fcdo]	0 [fcgl]
near closure (nc)	—	0 [ncld]	0 [ncal]	0 [ncpo]	0 [ncpa]	0 [ncve]	0 [ncgl]
ac/lc/vi	0 [acla]	—	0 [lcap]	—	1 [acpa]	0 [viuv]	—

Tabelle 2 - Kodierung des Gestentyps am Beispiel der Geste j aus der in Abbildung 4 dargestellten gestischen Partitur. Die ersten beiden Buchstaben stehen für die Art des Verschlusses (ac, lc stehen für approximant und lateral closure und vi für vibrant). Die letzten beiden Buchstaben stehen für den Artikulationsort.

Vokalische Gestenlabel	vorne ungerundet	halbvorne ungerundet	vorne gerundet	halbvorne gerundet	halbhinten gerundet	hinten gerundet
hoch	0 [i]	0 [I]	0 [y]	0 [Y]	0 [U]	0 [u]
halbhoch	0 [e]	—	0 [2]	0 [9]	—	0 [o]
halbtief	1 [E]	0 [@]	—	—	0 [O]	—
tief	0 [a]	0 [6]	—	—	—	—

Tabelle 3 - Kodierung des Gestentyps am Beispiel der Geste E, wie sie in der gestischen Partitur aus Abbildung 4 vorkommt. Die Buchstaben in den eckigen Klammern sind Label für die unterschiedlichen vokalischen Gestentypen. Sie bezeichnen Ziele für die Artikulatorbewegung und sind keine Phoneme.

definiert, die das Gestenlabel entsprechend der Tabelle 2 kodieren, und für die Gruppen VS1 und VS2 wurden Neuronen definiert, die das Gestenlabel entsprechend der Tabelle 3 kodieren.

2.2.4 Ausrichtung der auditiven neuronalen Repräsentation und des Motorplans

Für die zeitliche Vergleichbarkeit der unterschiedlichen Items wurden sie an einem Bezugspunkt ausgerichtet. Als Bezugspunkt wählte man den Beginn des Silbenkern. Hier ist er nach [5] definiert als das Ende der Targetphase der Geste, die zum letzten Konsonant im Silbenonset gehört. Bei der Generierung der neuronalen Repräsentationen wurde der Beginn des Silbenkerns an den Zeitpunkt 225 ms gesetzt.

2.3 Training der Selbstorganisierenden Karte

Eine SOM besteht aus einer Eingabeschicht und einer Kohonenschicht. Alle Neuronen aus der Kohonenschicht (Kohonenneuronen) sind mit allen Neuronen aus der Eingabeschicht (Eingabeneuronen) verbunden. Jedem dieser Verbindungen ist ein Verbindungsgewicht zugewiesen. In diesen Verbindungsgewichten, die hier Werte zwischen 0 und 1 annehmen können, werden beim Training die Items gespeichert. Vor dem Trainingsvorgang werden die Verbindungsgewichte mit Zufallswerten zwischen 0 und 1 initialisiert. Während des Trainings werden dem SOM hintereinander die Trainingsitems präsentiert. Ein Trainingsschritt besteht dann aus der Präsentation eines Items und der Modifizierung der Verbindungsgewichte entsprechend des präsentierten Items. Nach dem Training lässt sich aus den Verbindungsgewichten eine bestimmte Ordnung der erlernten Items extrahieren [3].

Um nun die phonemische, auditive und motorische Repräsentation eines Items miteinander zu assoziieren, müssen diese drei Repräsentationen der SOM gleichzeitig präsentiert werden. Nach der in Abschnitt 2.2 beschriebenen neuronalen Enkodierung dieser Repräsentationen erhält man eine neuronale Darstellung eines Items mit $171 + 1560 + 983 = 2714$ Neuronen. Daher wurde

eine SOM mit 2714 Eingabeneuronen und mit 625 Kohonenneuronen definiert. Vesuche haben gezeigt, dass 625 Neuronen, quadratisch angeordnet (25×25), ausreichten um die Items zu erlernen. Die Trainingsmenge besteht mit Wiederholungen aus 731 Items. Diese Items wurden der SOM in 100 Trainingsdurchläufen als Eingabe präsentiert, dabei wurde die Reihenfolge der Items bei jedem Durchlauf randomisiert. Für die Veränderung der Verbindungsgewichte in jedem Trainingsschritt muss zuerst das Kohonenneuron gefunden werden, dessen Verbindungsgewichte zu den Werten der Eingabe am ähnlichsten ist. Dieses Neuron wird Gewinnerneuron genannt. Der Gewinnerneuron wird bestimmt, indem man die Eingabewerte und die Verbindungsgewichte eines Kohonenneurons als Vektoren auffasst und die Euklidische Distanz zwischen ihnen berechnet. Das Kohonenneuron mit der geringsten Distanz ist der Gewinnerneuron win_t. Die Verbindungsgewichte des Gewinnerneurons $c_{win}(t)$ und die Verbindungsgewichte $c_n(t)$ der Neuronen in der Nachbarschaft $N(t)$ des Gewinnerneurons werden nach

$$c_{win}(t+1) = c_{win}(t) + \alpha(t) \cdot [c_{win}(t) - input(t)] \quad (1)$$
$$c_n(t+1) = c_n(t) + \alpha(t) \cdot [c_n(t) - input(t)] \quad (2)$$

verändert, wobei $input(t)$ die vektorielle Darstellung der Eingabe und $\alpha(t) \in [0,1]$ die Lernrate im Trainingsschritt t ist. Hier wurde eine Anfangslernrate von $\alpha(0) = 0{,}8$ gewählt. Die Nachbarschaft eines Kohonenneurons ergibt sich aus der quadratischen Anordnung der Kohonenschicht. Eine Nachbarschaft von 1 um ein bestimmtes Neuron beinhaltet alle 8 Neuronen, die um den betreffenden Neuron herum liegen, eine Nachbarschaft von 2 die 24 umliegenden Neuronen usw.. Als Anfangsnachbarschaft wird $N(0) = 15$ gewählt. Es ist notwendig eine hohe Anfangsnachbarschaft $N(0)$ zu wählen, damit auf der SOM zusammenhängende Regionen entstehen. Ansonsten kann es passieren, dass für ein Trainingsitem Regionen an unterschiedlichen Stellen im SOM entstehen. Die Nachbarschaft und die Lernrate verringert sich mit jedem Trainingsschritt um einen Verfallsfaktor $a \in [0,1]$ nach

$$N(t+1) = N(t) \cdot a \quad (3)$$
$$\alpha(t+1) = \alpha(t) \cdot a. \quad (4)$$

Die Wahl des Verfallsfaktors a hängt von der Anzahl der Trainingsschritte insgesamt ab und wurde so gewählt, dass gegen Ende des Trainings die Lernrate und die Nachbarschaft gegen 0 konvergiert. Bei diesem Training wurde daher $a = 0{,}9999$ gesetzt.

3 Ergebnisse

In Abbildung 5 ist die Anordnung der Items nach einem Trainingsdurchlauf zu sehen. Die Quadrate repräsentieren die Kohonenschicht der SOM. Die Verbindungsgewichte der SOMs wurden entsprechend der in Abschnitt 2.2 vorgestellten neuronalen Kodierungen wieder dekodiert. Dabei wurde ein Neuron als phonetischer Repräsentant einer Silbe angenommen, wenn der Aktivierungsgrad zur phonologischen Repräsentation dieser Silbe höher als 0.8 (80 %) liegt. In jedes Quadrat wurde die phonemische Segmentfolge eingetragen, die sich aus der Dekodierung der Vebindungsgewichte des zugehörigen Neurons ergibt. Man sieht, dass ähnliche Segmentfolgen auch im SOM benachbart dargestellt werden (z. B. unten links: s@, S@, d@, p@ b@, t@, k@, g@ oder in der Mitte: '?Im, '?I, '?In, 'fIn, 'kIn). Abbildung 6 zeigt den Fortschritt des Trainings an. Die Anzahl der Neuronen, die ein Item repräsentierten, waren in der Anfangsphase des Trainings und in der Endphase des Trainings annähernd gleich, während sie in der zwischendurch zurückgegangen ist (Abb. 6 oben). Die Anzahl der in der SOM gespeicherten Items sind mit fortschreitendem Training gestiegen. Nach einem Training mit 100 Trainingsdurchläufen

’?IC	’?IC	’zi:t		’mIt	’mIt		g@n	g@n		’mi6		’zi:	’zi:	’zi:	’?i6	’?i6	’ja	’ja	d6	d6		’vi6		’da:
			t@t		’mIt	’nOx	g@n		ma		’hOI		’zi:	’?i:	’?i6	’?i6					’vE6		’baI	
’zIC		’mUs		nOk	nO	’nOx		n@m		n6		ma:		’?i:	’?i:	’nICt	’vaI	’vaI		’bE6		’ze:		jo:
					m@n	’nOx	N@n	n@m	’ha:	n6	’ra:	ma:		’?i:	’nICt	’nICt	’nICt		baI		’bE:			
’?Es	’?Es			m@n	m@n		N@n		’ha:	’ha:	’ra:	’ra:	’kOn	’kOn	’nICt	’nICt				’?aUx				’va6
’?Es											’vi:	’kOn	’kOn	’kOn	’?E6		t6		’da	’?a	’?a			’va6
	’bIs		’vas		’?Ist		’ma			’vi:	’vi:	’vi:		’?E6	’?E6		t6	t6		’?a:		’gants		’laN
was		l@		r@	’?Ist			’le:			’vi:		m@	’?aI	’?E6	’?o:	t6	’?am	’?a	’?a:			’ma:l	
’das	li	l@	r@	r@	r@	z@	z@			’li:		N@		’?aI		’?o:		’?am	’?am	’klaI	’mo:nt	’mo:nt		’na:x
’das	’das	’hat					z@	’zo:			’ri:		n@	’?aI	’?aUf	’?o:		’?am	’klaI	’klaI	’klaI	’mo:nt		’na:x
’das	’hat	’hat		’?an				’zo:		di:		n@	n@		’?aUf	’?aUf	’dE6	tsIm	tsIm				’ve:k	
’jEtst	’jEtst									’?In		’?Um	n@	’?aUs	’?aUf	’dE6	’dE6	’dE6		’tsvaI		’kaI		’me6
	’?alts	’?alts		’?Un		’?En		’?Im	’?I	’?In		’?Um		’?aUs	’?aUs	’fo6	’dE6	’fy6						
lIC		’?alts			’?Ins					’?In						’fo6	’f6	’fy6		ts@n		x@n		r@n
						’?ap		’kIn		’fIn		’zaI					f6	f6						
g@	g@		’bEt			’di:		’kIn								’So:n		fE6		f@n		z@n		k@n
	k@	k@		’dOx	’di:	’di:	’di:	’kOm		’gIpt		’tsUm		C@n	C@n		’fe:	fE6	g@l	f@n			k@l	k@n
t@	t@			’dOx	’de:m	’di:	’kan	’kOm	’kOmt		b6	’tsUm	’de:n	C@n		C@		g@l	g@l	g@l	’?Unt		b@n	
t@	b@	b@	b@	’de:m	’de:m		’kan	’kan	n@n	b6	b6		’de:n	’de:n	t@l		’fa:		g@l	’?Unt	’?Unt	’?Unt	b@n	b@n
pa:		b@		’vo:			’ka	n@n	n@n	n@n	b6		’de:n		t@l	’zOn		’vOl		dEl	’?Unt	’man	b@n	’vI6t
	p@		’?y:		’du:		’kat	’kat		g6	g6		’dEn		’zOn	’zOn	’zOn		’vIl			’man		’vI6t
’pa	p@	d@		hE				m6	m6	m6			’dan	’zaIn	’za	’zOn		vEn						
’pa		d@	d@		’tE				m6	l6	l6		’dan		’za:k		d@n			ni:		l@n		’?al
			S@	S@		’fraU		’SnEl	’SnEl	ts@		’tsu:				’haUs			’?i:n				’?an	
pi:		s@	s@		’fi:		’Spi:		ts@	ts@	ts@	’tsu:	tsu:			’ha				’?i:m	’?aIn	’?aIn	’?aIn	h@n

Abbildung 5 - Ordnung der Selbstorganisierenden Karte nach Training am Beispiel der phonemischen Repräsentation.

waren 90 % der Trainingsitems in der Selbstorganisierenden Karte enthalten (Abb. 6 unten). Auch die Eigenschaft der kortikalen Plastizität lässt sich in der SOM beobachten. Die Items, die dem Netz am häufigsten präsentiert wurden, belegen auch die größten Regionen in der SOM. Abbildung 7 zeigt die Größe einer zu einem Trainingsitem gehörenden Region im SOM in Abhängigkeit von dem Vorkommen des Trainingsitems in der Trainingsmenge.

4 Diskussion

Das Modell MSYL ist eine erste Realisierung eines mentalen Silbenspeichers. Nach dem Training ergibt sich eine Topographie mit Regionen, deren Ausdehnung von der Häufigkeit eines Items in der Trainingsmenge abhängt. Die Eigenschaft der kortikalen Pastizität als Grundprinzip selbstorganisierenden Lernens ist gegeben. In diesem Fall wurden die somatosensorischen Daten weggelassen, da sie im Vergleich zum Motorplan keine zusätzlichen Informationen lieferten. In weiterer Betrachtung könnten somatosensorische Information ins Training mit einbezogen werden. Genauso könnte man zusätzlich Lippenrundung und Lippenverschluss als visuelle Information in das Training mit einbeziehen.

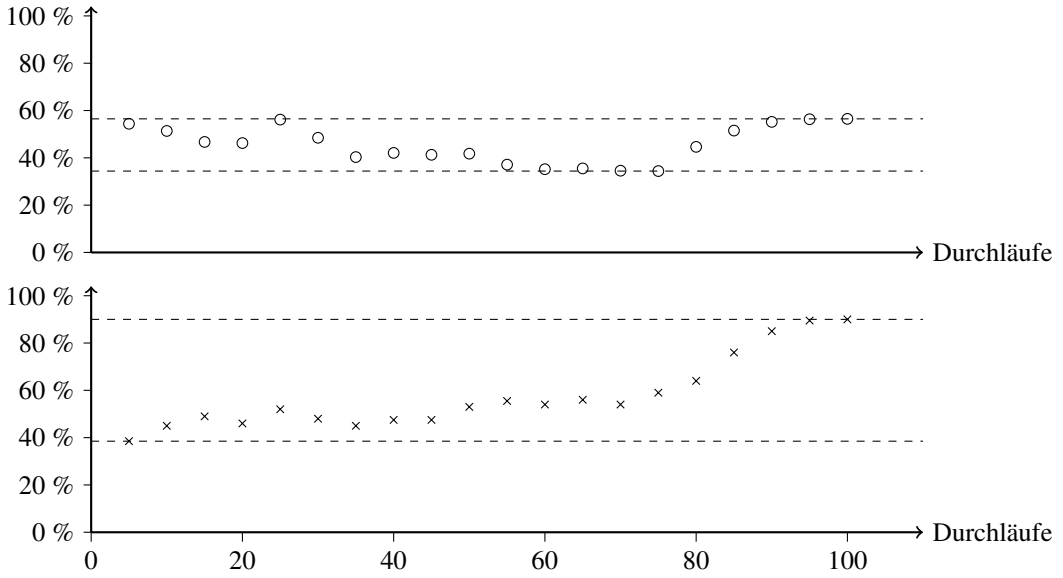

Abbildung 6 - Oben: Anteil der Neuronen der Selbstorganisierenden Karte, die einer Silbe zugeordnet werden können. Unten: Anteil der von der Selbstorganisierenden Karte repräsentierten Silben.

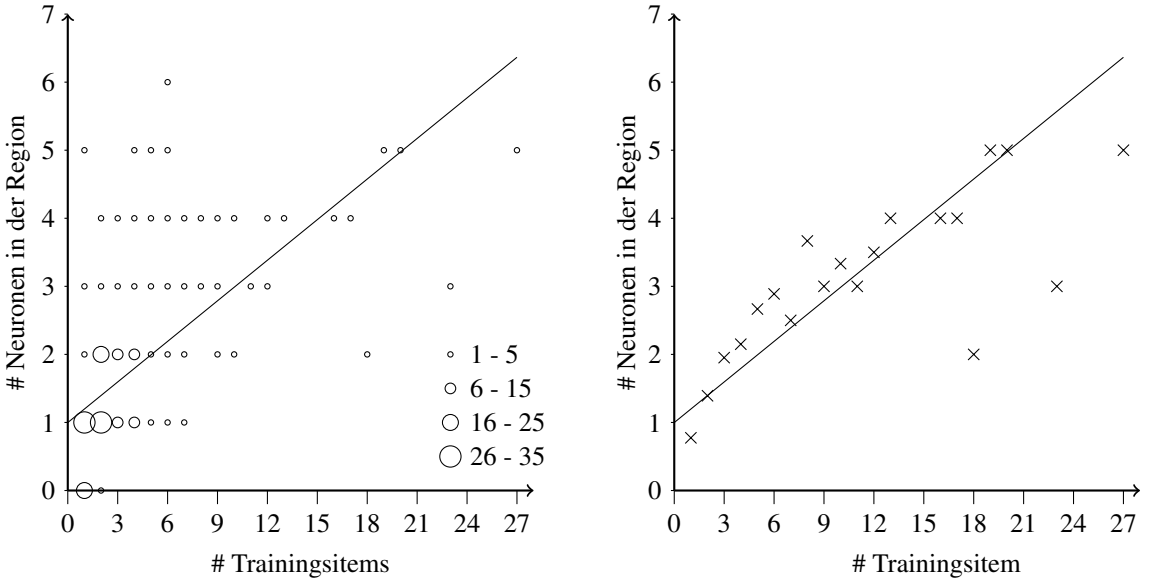

Abbildung 7 - Kortikale Plastizität: Die Abszisse repräsentiert die Anzahl der Vorkommen eines Items in der Trainingsmenge. Die Ordinate stellt die Anzahl der Neuronen der Region im SOM dar, zu dem dieses Item gehört. Links sieht man die Daten für jedes Item aufgetragen. Der Durchmesser der Kreise vergrößert sich mit zunehmender Anzahl der Punkte die übereinander liegen (siehe Legende). Rechts wurden zuerst alle Items gesammelt, die in der Trainingsmenge dieselbe Häufigkeit haben und anschließend der arithmetische Mittelwert über die Anzahl der Neuronen im SOM dargestellt, die in der Region zu diesem Item liegen.

Literatur

[1] BIRKHOLZ, P.: *3D-Artikulatorische Sprachsynthese*. Doktorarbeit, Universität Rostock, 2005.

[2] GUENTHER, F. H., S. S. GHOSH und J. A. TOURVILLE: *Neural modelling and imaging of the cortical interactions underlying syllable production*. Brain & Language, 96:280–301, 2006.

[3] KOHONEN, T.: *Self-organizing maps*. Springer, 2001.

[4] KRÖGER, B. J. und P. BIRKHOLZ: *A gesture-based concept for speech movement control in articulatory speech synthesis*. In: ESPOSITO, A., M. FAUNDEZ-ZANUY, E. KELLER und M. MARINARO (Hrsg.): *Verbal and Nonverbal Communication Behaviours*, S. 174 – 189. Springer Verlag, Berlin, Heidelberg, 2007.

[5] KRÖGER, B. J., P. BIRKHOLZ, J. KANNAMPUZHA, E. KAUFMANN und C. NEUSCHAEFER-RUBE: *Towards the Acquisition of a Sensorimotor Vocal Tract Action Repository within a Neural Model of Speech Processing*. 2011.

[6] KRÖGER, B. J., J. KANNAMPUZHA und C. NEUSCHAEFER-RUBE: *Towards a neurocomputational model of speech production and perception*. Speech Communication, 51:793 – 809, 2009.

[7] LEVELT, W. J. M.: *Spoken Word Production: A Theory of Lexical Access*. Proceedings of the National Academy of Sciences of the USA.

[8] WELLS, J. C.: *SAMPA computer readable phonetic alphabet*. In: GIBBON, D., R. MOORE und R. WINSKI (Hrsg.): *Handbook of Standards and Resources for Spoken Language Systems*, Kap. IV-B. Mouton De Gruyter, Berlin and New York, 1997. http://www.phon.ucl.ac.uk/home/sampa.