Growing SOM’s in Biologically Inspired Models of Speech Processing

Bernd J. Kröger\textsuperscript{1,2} & Mengxue Cao\textsuperscript{3}

\textsuperscript{1}Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany

\textsuperscript{2}Cognitive Computation and Applications Laboratory, School of Computer Science and Technology, Tianjin University, Tianjin, China

\textsuperscript{3}Laboratory of Phonetics and Speech Sciences, Institute of Linguistics, Chinese Academy of Social Sciences (CASS), Beijing, China
Outline

- Introduction
- SOM/GSOM models
  - Semantic level
  - Phonetic level
- Perspective

Interconnected growing self-organizing maps for auditory and semantic acquisition modeling

Mengxue Cao¹*, Aijun Li¹, Qiang Fang¹, Emily Kaufmann² and Bernd J. Kröger³, ⁴

¹ Laboratory of Phonetics and Speech Science, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
² Department of Special Education, Faculty of Human Sciences, University of Cologne, Cologne, Germany
³ Neurophotonics Group, Department of Phoniatrics, Pediaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany
⁴ Cognitive Computation and Applications Laboratory, School of Computer Science and Technology, Tianjin University, Tianjin, China
Outline

• Introduction

• SOM/GSOM models
  – Semantic level
  – Phonetic level

• Perspective

Interconnected growing self-organizing maps for auditory and semantic acquisition modeling

Mengxue Cao¹*, Aijun Li¹, Qiang Fang¹, Emily Kaufmann² and Bernd J. Kröger³,⁴

¹ Laboratory of Phonetics and Speech Science, Institute of Linguistics, Chinese Academy of Social Sciences, Beijing, China
² Department of Special Education, Faculty of Human Sciences, University of Cologne, Cologne, Germany
³ Neurophonetics Group, Department of Phoniatrics, Pedaudiology, and Communication Disorders, Medical School, RWTH Aachen University, Aachen, Germany
⁴ Cognitive Computation and Applications Laboratory, School of Computer Science and Technology, Tianjin University, Tianjin, China
Goal

- **Neurobiologically inspired quantitative** (computer-implemented) **model** of
  speech acquisition, production, perception

- **SOMs / GSOMs** use **brain-related learning principles**:
  - associative learning in our architecture
  - leads to self-organization and adaptation
  - leads to topological ordering of items -&gt; biological basis of phonetic features / semantic features
  - simple Hebbian learning, not: spike-timing-dependent plasticity (STDP)

- Become more realistic: e.g. use **NENGO** (neural simulation package -&gt; see my poster!)
  - includes neural noise generation; e.g. using leaky-integrate-and-fire neurons (LIF neurons)
  - time progression in 1 msec intervals (not 10-20 msec intervals)
learning process / structure:

- Baby/toddler/child dominates the face-to-face-communication scenario!
- Resulting maps and mappings are basis for developing phonological awareness!
- Babbling (sensori-motor-mapping)

The Model

- Mental lexicon
- S-MAP GSOM
- Neural mapping
- Semantic representation
- Feature vectors
- Phonemic processing
- Auditory external processing
- Auditory internal processing
- Somatosensory processing
- GSOM neural mapping
- Feature vectors
- Feedback loop
- Triangulation
- Self-perception

Imitation: (language-specific)

Caretaker

Interconnected growing self-organizing maps for auditory and semantic acquisition modeling

Mengxue Cao, Aijun Li, Qiang Fang, Emily Kauffmann and Bernd J. Kröger
Outline

• Introduction
• SOM/GSOM models
  – Semantic level: GSOMs
  – Phonetic level
• Perspective
Learning Data (Training Corpus)

- **Standard German Children’s Book Corpus**: phonetic transcriptions of:
  - 40 children’s books, targeted for 1 to 6 years old children
  - In total: 6513 sentences; 70512 words
  - 332 words with frequency of occurrence above 10 times

<table>
<thead>
<tr>
<th>Frequency counts in corpus</th>
<th>Frequency counts in training data</th>
<th>German words</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>392</td>
<td>78</td>
<td>Mama</td>
<td>mom</td>
</tr>
<tr>
<td>278</td>
<td>55</td>
<td>Bär</td>
<td>bear</td>
</tr>
<tr>
<td>235</td>
<td>47</td>
<td>Papa</td>
<td>dad</td>
</tr>
<tr>
<td>217</td>
<td>43</td>
<td>Mond</td>
<td>moon</td>
</tr>
<tr>
<td>190</td>
<td>38</td>
<td>Kinder</td>
<td>children</td>
</tr>
<tr>
<td>147</td>
<td>29</td>
<td>Katze</td>
<td>cat</td>
</tr>
<tr>
<td>145</td>
<td>29</td>
<td>Frau</td>
<td>wife</td>
</tr>
<tr>
<td>106</td>
<td>21</td>
<td>Bett</td>
<td>bed</td>
</tr>
<tr>
<td>105</td>
<td>21</td>
<td>Mädchen</td>
<td>girl</td>
</tr>
<tr>
<td>104</td>
<td>20</td>
<td>Wasser</td>
<td>water</td>
</tr>
</tbody>
</table>

Item 332: 10 → 2
Training Corpus: Generation of semantic features

- Association procedure by 2 native speakers of Standard German: leads to **1715 semantic features** for 332 most frequent words.
- Dropping off features, occurring once only, but keeping words distinguishable → **724 features** (neural representation: “0” or “1”).

<table>
<thead>
<tr>
<th>Frequency</th>
<th>German features</th>
<th>English translation</th>
</tr>
</thead>
<tbody>
<tr>
<td>81</td>
<td>ist ein Gegenstand</td>
<td>is an object</td>
</tr>
<tr>
<td>63</td>
<td>hat zwei Augen</td>
<td>has two eyes</td>
</tr>
<tr>
<td>60</td>
<td>es gibt verschiedene Arten</td>
<td>there are different types</td>
</tr>
<tr>
<td>56</td>
<td>hat einen Kopf</td>
<td>has a head</td>
</tr>
<tr>
<td>47</td>
<td>hat eine Nase</td>
<td>has a nose</td>
</tr>
<tr>
<td>46</td>
<td>hat zwei Beine</td>
<td>has two legs</td>
</tr>
<tr>
<td>42</td>
<td>hat eine Haut</td>
<td>has a skin</td>
</tr>
<tr>
<td>37</td>
<td>ist aus Kunststoff</td>
<td>is made of plastic</td>
</tr>
<tr>
<td>37</td>
<td>ist aus Metall</td>
<td>is made of metal</td>
</tr>
<tr>
<td>36</td>
<td>ist ein Mensch</td>
<td>is a human</td>
</tr>
</tbody>
</table>

The table lists the ten most frequent semantic features.
after first training: mathematical background for GSOMs, see CAO et al. 2013

... displays semantic feature regions \(\rightarrow\) topology

“unsolved node” for 54 words

Clustering of words with respect to specific semantic categories
after first reinforcement training: non compact shape
after 10 reinf./review. trainings: compact network again
GSOM Training procedure

- Focusing on **S-Map** only in this study
- **first basic growing training**: step 1 → problem: unresolved nodes
- **ten cyclic checking processes** for reinforcement trainings; followed by reviewing trainings: step 2 to 21

![Graph showing the number of words represented by a node in the network](image)

Number of Word Representations

Maximum Number of Words Represented by a Node in the Network

number of “unsolved” words
GSOM Training procedure

Blue: total number of nodes
red: number of nodes, representing a word
→ 84% of all words at the end
(279 words at 934 neurons)

→ “empty nodes” are needed because of projection of all feature dimensions in 2D plane
Outline

• Introduction
• SOM/GSOM models
  – Semantic level
  – Phonetic level: SOMs
• Perspective
... Includes a Physiological Articulatory Model:

Dang & Honda (2004) JASA

- Finite element model
- 15 muscles: tongue (intrinsic & extrinsic), tongue tip, lower jaw
- control by muscle activation patterns (primary motor level)
The Neural Representations (Vowels)

- High level motor
  - Jaw (x, y)
  - Tongue tip (x, y)
  - Tongue dorsum (x, y)
  - Lip tube (length, area)

- Low level motor
  - Muscle activation pattern
    - Gg, gap, lg, vert, trans, gh, jawOp, mhyoid

- Auditory
  - Auditory signal
    - First three bark-scaled formants

- Somatosensory
  - Somatosensory signal
    - Contact and area
    - Midsagittal distance d
Testing two models:

- Hypothesis 1: there is a hierarchy! (see our old model)
  Only high-level representations (motor & sensory) are learned by one SOM (P-Map)
  **We need a further SOM (or EP-map) in order to calculate low-level motor representations form high-level ones**

- Hypothesis 2: no hierarchy:
  high- and low-level motor representation is learned by one sensorimotor SOM (P-Map) together with (high-level) sensory representations
Babbling: Generation of Training Items

- Generation of proto-vocalic states (training items) starts at high motor level: contour point level (tongue dorsum point);
- Interpolating between three cardinal vowels: high-front, high-back, low → 1146 training items
Generation of Training Items

- Use EP-map for generation of muscle activation patterns (low-level motor states);
- Different solutions are possible (one-to-many problem); here using a minimum energy principle (for muscle force/energy)
  - Example: three solutions of each of the three cardinal vowels
  - Minimum energy solution means: muscle activation is divided between articulators: into lower jaw and tongue movement (high jaw muscle activity e.g. in “high front” and “low”)

<table>
<thead>
<tr>
<th>TM</th>
<th>Muscle pairs</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>low-back</td>
<td>Ggm</td>
<td>ggp</td>
<td>hg</td>
<td>lg</td>
<td>tran-lg</td>
<td>jawOp</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0</td>
<td>0.9</td>
<td>0</td>
<td>0</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>0.2</td>
<td>0</td>
<td>3.1</td>
<td>0</td>
<td>2.0</td>
<td>5.7</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>3.4</td>
<td>3.5</td>
<td>4.9</td>
<td>0</td>
<td>6.1</td>
</tr>
<tr>
<td>high-front</td>
<td>Ggp</td>
<td>sg</td>
<td>lg</td>
<td>mh</td>
<td>tran-lg</td>
<td>jawCl</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.3</td>
<td>0</td>
<td>0.9</td>
<td>1.0</td>
<td>2.5</td>
</tr>
<tr>
<td></td>
<td>1.5</td>
<td>0.3</td>
<td>1.4</td>
<td>0</td>
<td>0</td>
<td>1.7</td>
</tr>
<tr>
<td></td>
<td>0.4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.2</td>
</tr>
<tr>
<td>high-back</td>
<td>Ggp</td>
<td>hg</td>
<td>sg</td>
<td>Lg</td>
<td>tran-lg</td>
<td>jawOp</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0.1</td>
<td>0.6</td>
<td>0</td>
<td>0.4</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>0</td>
<td>0.4</td>
<td>0.9</td>
<td>0</td>
<td>0.9</td>
</tr>
<tr>
<td></td>
<td>0.7</td>
<td>0</td>
<td>0.4</td>
<td>3.5</td>
<td>0</td>
<td>1.0</td>
</tr>
</tbody>
</table>
SOM Training for Both Model Hypotheses

• Show: sensorimotor state representation for 20x20 SOM; Hyp 1

Tongue dorsum contour point & formant pattern (Bark scaled)

→ ordering with respect to phonetic features!
SOM Training for Both Model Hypotheses

• Show: sensorimotor state representation for 20x20 SOM; Hyp 2

Tongue dorsum contour point & muscle activation pattern & formant pattern (Bark scaled)

→ **No** ordering with respect to phonetic features!
→ **No** learning of phonetic features! (high-level features!!)
→ Muscle force information is differently organized in comparison to high-level sensory and motor information
→ seems to “destroy” or “complicates learning” of high-level sensorimotor features
Motor-to-Auditory Association

- Calculate over all 20x20 learned states; for five trainings of each model hypothesis;
- and compare with direct calculation of auditory states by using articulatory-acoustic model; calculate mean error over all 400 states;
- Result: Hypothesis 1 (hierarchical model) performs better! (and learns features!) → we will use hierarchical model

<table>
<thead>
<tr>
<th></th>
<th>F1</th>
<th>F2</th>
<th>F3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Hypothesis 1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infant 1</td>
<td>8.06%</td>
<td>5.61%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Infant 2</td>
<td>8.9%</td>
<td>5.94%</td>
<td>1.59%</td>
</tr>
<tr>
<td>Infant 3</td>
<td>9.35%</td>
<td>5.93%</td>
<td>1.51%</td>
</tr>
<tr>
<td>Infant 4</td>
<td>8.84%</td>
<td>5.96%</td>
<td>1.56%</td>
</tr>
<tr>
<td>Infant 5</td>
<td>8.29%</td>
<td>5.71%</td>
<td>5.57%</td>
</tr>
<tr>
<td><strong>Hypothesis 2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Infant 1</td>
<td>17.18%</td>
<td>12.86%</td>
<td>1.7%</td>
</tr>
<tr>
<td>Infant 2</td>
<td>23.35%</td>
<td>11.55%</td>
<td>1.57%</td>
</tr>
<tr>
<td>Infant 3</td>
<td>9.75%</td>
<td>6%</td>
<td>1.63%</td>
</tr>
<tr>
<td>Infant 4</td>
<td>9.3%</td>
<td>6.23%</td>
<td>1.6%</td>
</tr>
<tr>
<td>Infant 5</td>
<td>14.25%</td>
<td>8.68%</td>
<td>1.63%</td>
</tr>
</tbody>
</table>
Resulting Model Structure

- Hypothesis 1 is convincing: we need a hierarchy!
- Only high-level representations (motor & sensory) are learned by one SOM (P-Map)
- We need a further SOM (or EP-map) in order to calculate low-level motor representations form high-level ones
Outline

• Introduction
• Rate models: Topography preservation
  – Semantic level
  – Phonetic level
• Perspective
Perspective

• SOM and GSOM models can be seen as biologically inspired modes
  – self organization leads to
  – topological ordering of items -> phonetic features, semantic categories

• Further work: Become more realistic: e.g. use NENGO (neural simulation package → Poster!)
  – includes neural noise generation; e.g. using leaky-integrate-and-fire neurons
  – includes time!
Many Thanks for Your Attention !!!

Mengxue Cao,
Chinese Academy of Social Sciences,
CASS, Beijing, China

References:

www search: “Bernd Kroeger homepage”
www.speechtrainer.eu