Statistical learning (Saffran, Aslin & Newport, 1996)

Presented 8 mo infants with stream of auditory syllables composed of four 3-syllable “words” in random order (with no word boundary cues)

\[
\begin{array}{cccc}
\text{Word 1} & \text{Word 2} & \text{Word 3} & \text{Word 4} \\
\ldots \text{pa bi ku} & \text{go la tu} & \text{da ro pi} & \text{ti bu do} \ldots \\
\text{Test Word} & \text{Test Part-word} \\
\end{array}
\]

\[\text{pabiku\ldots}\]

Giroux and Rey (2009)

- Auditory statistical learning (14 syllables, denoted by letters here)
- Input stream composed of randomly ordered disyllabic and trisyllabic “words”:
  \[\text{ABC DEF GH IJ KL MN} \]
  \[\text{IJKLMNOPQRSTU} \ldots\]
- Exposure for 2 min (400 syllables) or 10 min (2000 syllables)
- Compared preference for disyllabic words (GH IJ KL MN) or embedded pairs within trisyllabic words (AB BC DE EF), each against partwords (transitions across word boundaries; CD FG HI JK LM …)
- Note that disyllabic words and embedded pairs are matched on element frequency, pairwise frequency, transition probability, and conditional probability

Giroux and Rey (2009): Results

- Participants
  - SRN
  - Also tested an SRN trained on syllable prediction (IJDEF…: I⇒J, J⇒D, D⇒E, E⇒F, F⇒G…)

Statistical learning: Parts and wholes

- Statistical learning (Safran et al., 1996) is often cast as a means of discovering the \textbf{units} of perception (words, objects) through a process of \textbf{chunking} (Perruchet & Vinter, 1998; Thiessen et al., 2013)

- However, words and objects have \textbf{substructure}: parts that contribute systematically to the meaning/function of the whole
  - UN-TEACHEABLE
  - bicycle wheels, seat, handlebars, etc.

- What is the relation of the chunk for a whole and the chunks for its parts?
  - Typically considered to be alternative organizations

<table>
<thead>
<tr>
<th>Study</th>
<th>Words Mean (SE)</th>
<th>Part-words/Nonwords Mean (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Present study</td>
<td>6.78 (0.36)</td>
<td>7.36 (0.42)</td>
</tr>
<tr>
<td>Saffran, Aslin, and Newport</td>
<td>7.97 (0.41)</td>
<td>8.85 (0.45)</td>
</tr>
<tr>
<td>(1996), Experiment 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Saffran, Aslin, and Newport</td>
<td>6.77 (0.44)</td>
<td>7.60 (0.42)</td>
</tr>
<tr>
<td>(1996), Experiment 2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean listening time (SE)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Fiser and Aslin (2005) Experiment 1

- Presented series of 6-element displays constructed from 2 adjacent triples in 5x5 grid (11 min movie, 3s SOA, 1s ITI)
- Tested triples against random triples, and embedded pairs against random pairs

Fiser and Aslin (2005) Experiment 1: Results

- Clear preference for triples over random triples, but no preference for embedded pairs over random pairs

Embeddedness constraint

Fiser and Aslin (2005, p. 532):

Not all the embedded features that are parts of a larger whole are explicitly represented once a representation of the whole has been consolidated. We call this phenomenon the **embeddedness constraint** of statistical learning.... If there is a reliable mechanism that is biased to represent the largest chunks in the input in a minimally sufficient manner, rather than using a full representation of all possible features, this constraint can eliminate the curse of dimensionality.

**Problem:** Lots of evidence that parts and wholes work together

- Morphological priming (TEACHER ⇒ TEACH; Marslen-Wilson et al., 1994)
- Word superiority effect (E in READ vs. #E##; Reicher, 1969; Wheeler, 1970)
- Object superiority effect (noses in faces vs. alone; Tanaka & Farah, 1993)
- Lexical influenced on phoneme perception (Ganong, 1980)

Current approach

- Neural network learning of distributed representations of inputs
- Pressured to learn **efficient** representations due to limits on dimensionality (number of hidden units) and degree of nonlinearity (sigmoidal units; limited weight magnitudes)
- Will take advantage of shared structure (e.g., repeated subsets) and degree of **context dependence vs. independence**
  - Independence ⇒ “componential” representations (≈ small chunks)
  - Dependence ⇒ “conjunctive” representations (≈ large chunks)
- Relationship between different levels of structure (e.g., parts vs. wholes) depends on the structure of the domain
  - No “embeddedness constraint” but might behave so in some contexts
- No discrete “units” of perception; sensitivity to every level of structure is **matter of degree**
Simulation 1: Giroux and Rey (2009)

- Input stream composed of randomly ordered disyllabic and trisyllabic "words":
  ABC DEF GH IJ KL MN (IJDEFGHKLABCNMDEFGNHABCKLJGHKLABC....)
- SRN with 14 input units, 14 output units (1 per syllable for each), 30 hidden and context units
- Trained to predict next syllable: I → J, J → D, D → E, E → F, etc.
  - Same exposure as participants (400 or 2000 syllables)
- Testing: preference for words (real or embedded) over partwords
  \[
  1.0 - \frac{Error_{words}}{Error_{words} + Error_{partwords}}
  \]
- Tested either with or without prediction of final silence
  - Performance on AB (from ABC) should be worse because C is missing

Simulation 2: Fiser and Aslin (2005) Experiment 1

- Inputs: 12 units (1 per element) at each of 5x5 grid positions (300 total)
  - Noise added to unit activities at locations containing elements
- Autoencoder trained to reconstruct input over output via two layers of intermediate (hidden) units
  - 300 inputs → 80 hidden units → 40 hidden units → 300 outputs
- Same object configurations and testing comparisons as participants
- Positional invariance... Each configuration trained (and tested) at all 25 input positions (with wrap-around)
  - Removes any position-specific frequency differences
  - Precludes equating training exposure with that of participants
- Same network, parameters, training and testing procedures used for all remaining simulations

Simulation 1: Results

- Giroux & Rey SRN
- No prediction of silence
- Prediction of silence
- Giroux & Rey Participants

Simulation 2: Results

- Fiser & Aslin Expt. 1 results
- Simulation 2 results

Reconstruction of AB (in ABC) is poor because network also activates C

Very little variance in network performance
- Performance measure is based on relative mean error of relevant stimulus classes (not trial-by-trial 2AFC)

Can participants learn both triples and pairs simultaneously?

Simulation 3 results

Fiser & Aslin Expt. 3 results

Simulation 4: Fiser and Aslin (2005) Experiment 4

Quadruples and pairs; mid-exposure testing

Simulation 4 results

Simulation 5: Fiser and Aslin (2005) Experiment 5

Base-sextuple 1  Scene  Base-sextuple 2

Strong pairs  Weak pairs

Orbán, Fiser, Aslin, and Lengyel (2008)

Compared a Bayesian Chunk Learner (BCL) and an Associative Learner (AL) against human performance (including Fiser & Aslin, 2005, Expts. 1, 4)

BCL determines likelihoods of all possible chunk inventories given all training displays and priors on number and scales of chunks, then uses these to calculate posterior probabilities of test displays
  - Doesn’t actually assign an organization to a display, nor learn from each display incrementally

AL has same overall structure as BCL but learns only pairwise correlations between elements (no notion of “chunk”)

Both BCL and AL account for results of previous studies
  - AL fit to Fiser and Aslin (2005) Expt. 4 was poor
  - Didn’t address Fiser and Aslin (2005) Expt. 5

Carried out new study to dissociate predictions of BCL and AL (by matching pairwise correlations)

Six-element displays (triple + pair + singleton or quadruple + pair)

Compare “true” triples (green) to “false” triples (embedded in quadruple (red) but never occurring independently)

Orban et al. results

Simulation 6 results

Simulation 7: Joint effects of parts and wholes

Based on Sim. 4 (quads and pairs); one quad has “parts” that occur independently

Compare Whole + Parts to Whole-only (other quad) and Parts-only (other pairs) (Whole frequencies not matched)

Results

Whole + Parts ≈ Whole-only (2x) (benefit of parts)
Whole + Parts > Parts-only (benefit of whole)

Conclusions

“Units” of perception may not correspond to discrete entities
- The extent to which a particular subset of the input in a particular context is represented in a coherent manner is a matter of degree
- Whether “parts” and “wholes” cooperate or compete (or are even represented at all) is not stipulated in advance but arises naturally as a consequence of incidental learning in the domain
- Statistical learning is much richer than “chunking”

“Associative” learning: Task vs. mechanism
- Is learning based solely on surface features, or can the system learn to re-represent inputs so as to alter their relative similarities?
- Networks that learn associative tasks are not necessarily associative learners

Statistical learning is much richer than “chunking”