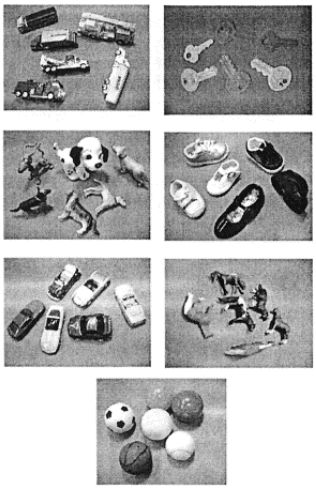
Learning words from sights and sounds: a computational model

Deb K. Roy , and Alex P. Pentland
Presented by Xiaoxu Wang

Introduction

- Infants understand their surroundings by using a combination of evolved innate structures and powerful learning abilities.
- They developed a computational model called Cross-channel Early Lexical Learning (CELL).
- It acquires words from multimodal sensory input and learns by statistically modeling the structure.

Infant-directed speech Experiments



- Participants were asked to engage in play centered around toy objects
- The infants could not produce single words.
- The caregivers reported varying levels of limited comprehension of words.

Object	Utterance
dog	He's gonna run and hide
dog	He's gonna hide behind my shoe
dog	Look, Savanah
dog	See his eyes?
dog	You like anything with eyes on it, eh?
dog	Just like you he has eyes
dog	Ruf ruf ruf
car	That's what your daddy likes, look!
car	Doors open vroom!
car	The seats go forward, and they go back!
shoe	You're always climbing into the shoes at home
shoe	Savanah! (infant's name)
truck	OK, you want it to drive?
truck	The wheels go around
truck	Your uncle Pat drives a truck like that
dog	He has a red collar
key	Let me see it
key	Do the keys have teeth?
key	You only have two teeth

Problems of early lexical acquisition

- Three questions of early lexical acquisition
  - Discover speech segments which correspond to the words of their language.
  - How to learn perceptually grounded semantic categories?
  - How to learn to associate linguistic units with appropriate semantic categories?

Speech Segmentation

- Let us do an experiment

  I am going to say three sentences in Chinese. Could you tell me how many words in the first sentence? What is the word corresponding to this object?

Speech Segmentation

- Let us do an experiment

  I am going to say three sentences in Chinese. Could you tell me how many words in the first sentence? What is the word corresponding to this object?
  - I am holding a pencil.
  - This is my pencil.
  - Pencils are useful.

Background

- Existing speech segmentation models may be divided into two classes
  - Based on local sound sequence patterns or statistics. The model was trained by giving it a lexicon of valid words of the language. To segment utterances, the model detect all trigrams which did not occur word internally during training. 37% word boundry detection
  - Minimum description length (MDL)

Spoken utterance

- Spoken utterance are represented as array of phoneme probabilities.
  - The acoustic input is put though a filter called Relative Spectral-Perceptual Linear Prediction (RASTA-PLP) . The filter is designed to attenuate nonspeech components of an acoustic signal. It does so by suppressing spectral components that change either faster or slower than the speech.

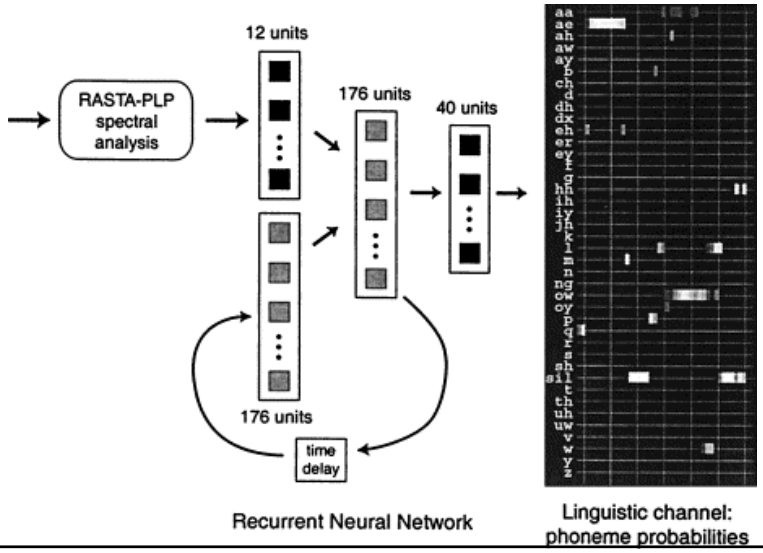
Spoken utterance

- Filtered signal is expanded using an exponential transformation and each power band is scaled to simulate laws of loudness perception in humans.
- A 12-parameter representation of the smoothed spectrum is estimated from a 20 ms window of input.
- The window is moved in time by 10 ms increments resulting in a set of 12 RASTA-PLP coefficients estimated at a rate of 100 Hz.

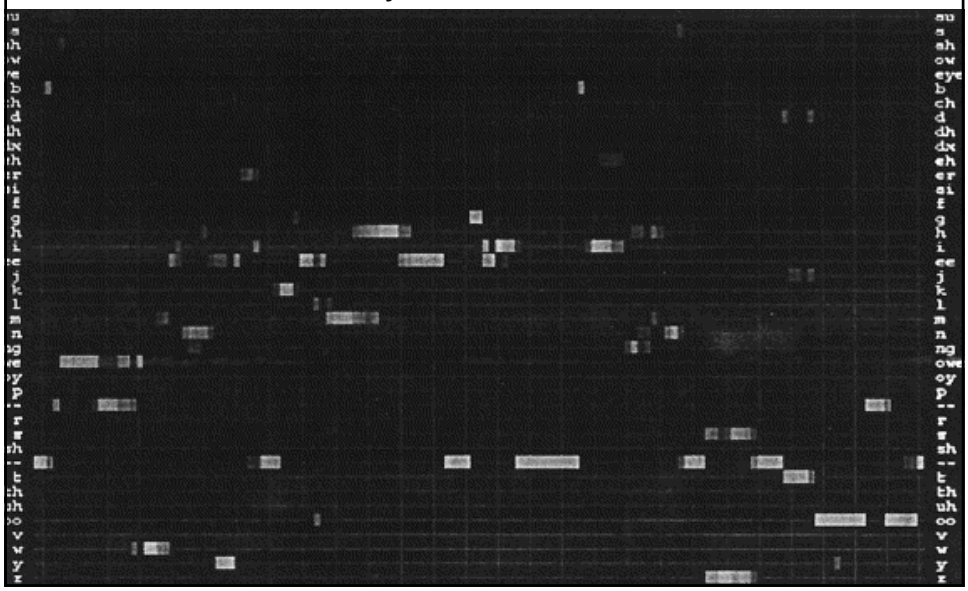
Recurrent Neural Network

- A recurrent neural network analyses RASTA-PLP coefficients to estimate phoneme and speech/silence probabilities.
  - The RNN has 12 input units, 176 hidden units, and 40 output units.
  - The 176 hidden units are connected through a time delay and concatenated with the RASTA-PLP input coefficients.
  - The time delay units give the network the capacity to remember aspects of old input and combine those representations with fresh data.

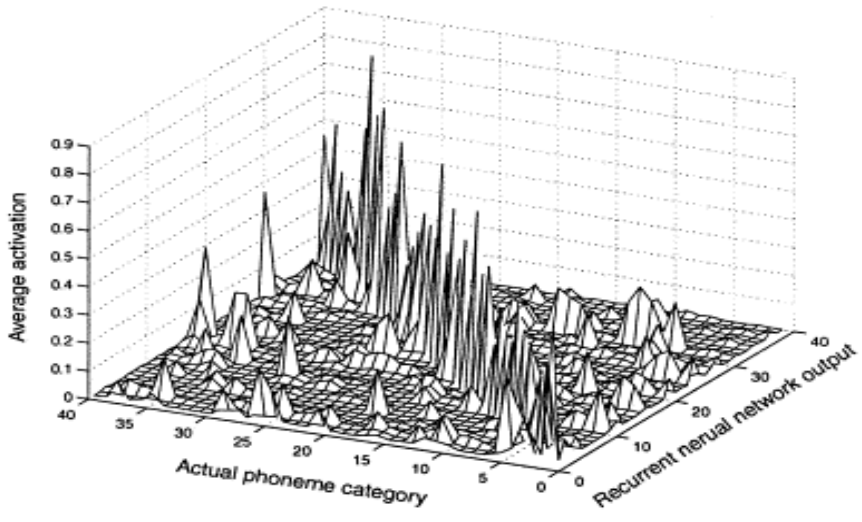
Recurrent Neural Network



Sample output from the recurrent neural network for the utterance "Oh, you can make it bounce too!"



The performance of the RNN



Speech Segmentation

- The RNN outputs are treated as state emission probabilities in a Hidden Markov Model (HMM) framework. The Viterbi dynamic programming search, is used to obtain the most likely phoneme sequence for a given phoneme probability array. The system obtains
  - The most likely sequence of phonemes which were concatenated to form the utterance
  - The location of each phoneme boundary for the sequence.

Speech Segmentation

- Any subsequence within an utterance terminated at phoneme boundaries is used to form word hypotheses.
- Additionally, any word candidate is required to contain at least one vowel. This constraint prevents the model from hypothesizing consonant clusters as word candidates. We refer to a segment containing at least one vowel as a *legal segment*.

Comparing words

- It is possible to treat the phoneme sequence of each speech segment as a string and use string comparison techniques.
- A limitation of this method is that it relies on only the single most likely phoneme sequence.
- A sequence of RNN output contains additional information which specifies the probability of all phonemes at each time instance. To make use of this additional information, they developed the following distance metric.

Comparing words

Two segments α_i and α_j can be decoded as phoneme sequences Q_i and Q_j . Q_i and Q_j can generate HMMs λ_i and λ_j. We wish to test if the hypothesis λ_i can generated α_j .

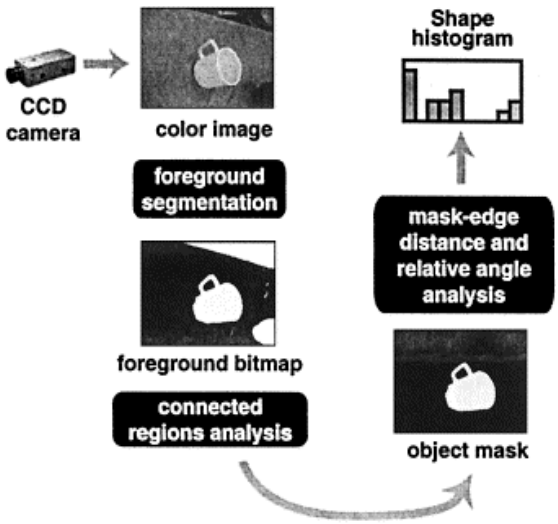
$$d_A(\alpha_i, \alpha_j) = -\frac{1}{2}\left\{\log\left[\frac{P(\alpha_i|\lambda_j)}{P(\alpha_i|\lambda_i)}\right] + \log\left[\frac{P(\alpha_j|\lambda_i)}{P(\alpha_j|\lambda_j)}\right]\right\}$$

Empirically, the result metric was found to return small values for words which humans would judge as phonetically similar.

Visual Input

- Similar to speech input, the ability to represent and compare shapes is also built into CELL.
- Three-dimensional objects are represented using a view-based approach in which two-dimensional images of an object captured from multiple viewpoints collectively form a visual model of the object.

Visual Input



Object shapes are represented in terms of histograms of features derived from the location of object edges.

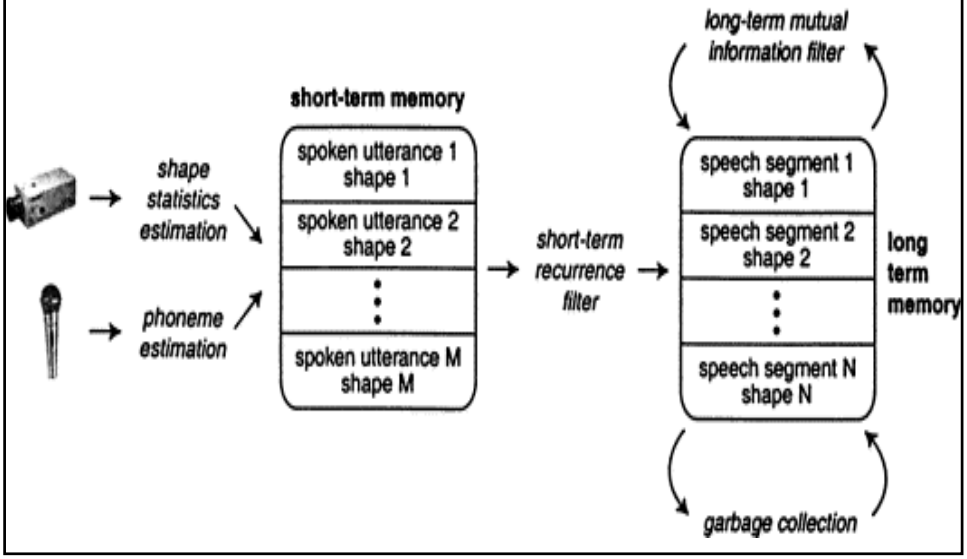
Comparing Visual Input

Using multidimensional histograms to represent object shapes allows for direct comparison of object models using information theoretic or statistical divergence functions. In practice, an effective metric for shape classification is the

. χ^2-divergence:

$$d_{\chi^2}(X, Y) = \chi^2(X, Y) = \sum_i \frac{(x_i - y_i)^2}{x_i + y_i}$$

The Structure of CELL



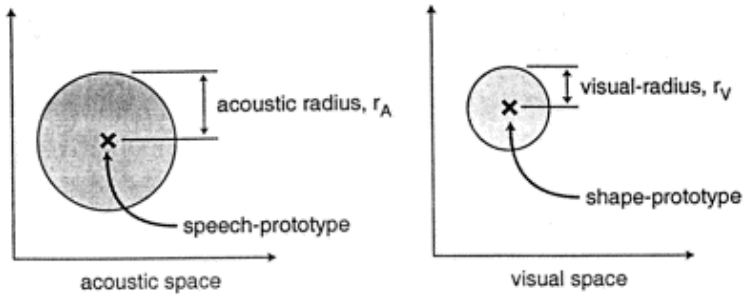
Word Learning

- Objctive: Each utterance may consist of one or more words. Similarly, each context may be an instance of many possible shape categories. Given a pool of utterance-context pairs, the learner must infer speech-to-shape mappings (lexical items) which best fit the data.
- Short term memory (STM), pass the pairs (prototypes) with high local recurrency to long term memory (LTM). For example,

dog-----  the-----  ball----- 

Word Learning

LTM create lexical items by consolidating AV-prototypes based on a mutual information criterion. This consolidation process identifies clusters of AV-prototypes which may be merged together to model consistent intermodal patterns across multiple observations.



Mutual Information

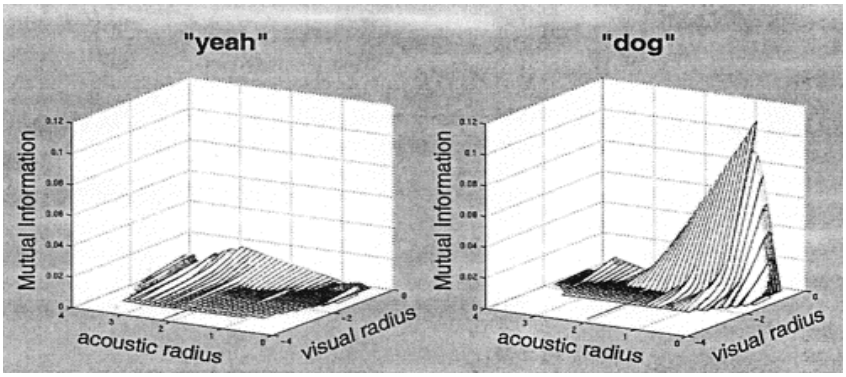
- A=1 iff distance_A(x, y) <=r_A, V is similar.
- The probabilities are estimated using relative frequencies of all *n* prototypes in LTM.

$$P(V = j) = \frac{|V = j|}{n} \qquad P(A = i) = \frac{|A = i|}{n}$$

$$P(A = i, V = j) = \frac{|A = i, V = j|}{n}$$

$$I(A; V) = \sum_i \sum_j P(A = i, V = j) \log \left[\frac{P(A = i, V = j)}{P(A = i)P(V = j)} \right]$$

Mapping



The prototype “yeah”-”dog” found little support from other AV-prototypes in LTM which is indicated by the low flat mutual information surface. In contrast, in the example on the right, the word "dog" was correctly paired with a dog shape.

Evaluation measures

- Lexical items obtained from speaker data sets are evaluated by
  - Segmentation accuracy
  - Word discovery

    dog Accepted /dg/, /g/ and /ðdg/ (*the dog*)

    Rejected /dgIz/ (*dog is*) .
  - Semantic accuracy

    The best choice of the meaning of a prototype is whatever context co-occurred with it.

Result

Rank	Phonetic transcript	Text transcript	Shape category	Segment.accuracy	Word Disc.	Semantic accuracy
1	ʃu	shoe	shoe E	1	1	1
2	fair ə	fire∗	truck D	0	1	1
3	rək	∗truck	truck C	0	1	1
4	dɑg	dog	dog D	1	1	1
5	Inəʃ	in the∗	shoe A	0	0	0
6	ki	key	key C	1	1	1
7	ki	key	key E	1	1	1
8	dɑggi	doggie	dog C	1	1	1
9	bɑl	ball	ball C	1	1	1
10	bɑl	ball	ball A	1	1	1
11	kiə	key∗	key C	0	1	1
12	ʌʃu	a shoe	shoe B	0	1	1
13	ənðIsIz	∗and this is	shoe B	0	0	0
14	(ono.)	(engine)	truck A	—	—	—
15	(ono.)	(barking)	dog A	—	—	—
Total				54%	85%	85%