Psychology 452

Week 6: Case Studies In Multilayer Perceptrons

Text-to-speech and neural networks

Metric spaces and neural networks

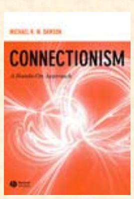
Nonmetric spaces and neural networks

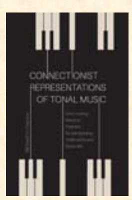
Course Trajectory

When	What
~~Weeks 1-3~~	~~Basics of three architectures (DAM, perceptron, MLP)~~
Weeks 4-6	**Cognitive science of DAMs and perceptrons**
Week 7	Connectionism and Cognitive Psychology
Weeks 8-10	Interpreting MLPs
Weeks 11-13	Case studies (interpretations, applications, architectures)

Discussion?

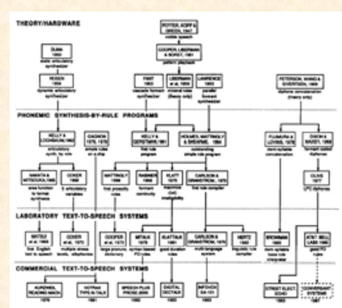
- **Questions, comments or issues?**





Case Study 1: Text To Speech

- **There is a long history of developing technologies for converting text to speech**
- **Klatt (1987) published an influential review that covered [several decades of research](#) in this domain**



Text To Speech Challenges

- Converting text to speech is a nontrivial task
- There are many irregular relationships between graphemes and phonemes
  - Ghti could be pronounced fish, from graphemes in enough and nation
- In general, text must be converted to various abstract linguistic representations in order to generate proper, realistic speech
- And the generation of realistic sounding speech is nontrivial too!

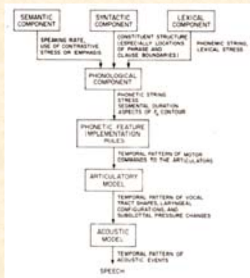
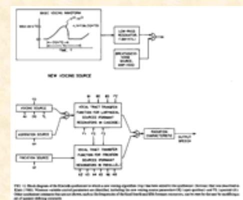


FIG. 2. Simplified block diagram of how a linguist might view the sentence generation process. An abstract linguistic representation for a sentence that is provided by the semantic component, syntactic component and lexical component undergoes various intermediate transformations before becoming an acoustic waveform.

Klattalk

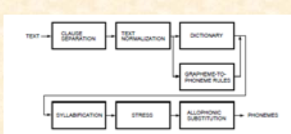
- Klatt's pioneering work at MIT produced a system that was first known as MITalk and later as Klattalk
- Klatt's interest was in exploring a diversity of representations of text, linguistic structures, and speech variables
- The synthesizer component of Klattalk alone required specifying the values of 19 different control parameters.


Dennis H. Klatt



DECtalk

- DEC developed Klattalk into a commercial system called DECtalk
- It was a computer peripheral that sold for about $4000 in the 1980s
- DECtalk looked text up in a dictionary (of irregular words) that converted it to sets of phonemes
- Otherwise, DECtalk used a set of grapheme-to-phoneme rules
- DECtalk's dictionary held 15,000 irregular words and it used more than 1,500 rules



DEC Talk Examples

- DECtalk was a prototypical classical system, and it was very successful
- DECtalk's dictionary held 15,000 irregular words and it used more than 1,500 rules
- It could be used to generate a number of different sounding voices
  - Examples of DECTalk voices

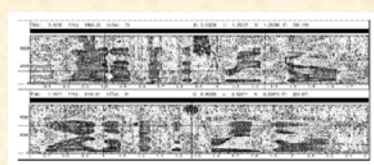
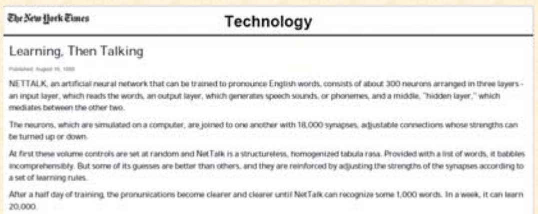


Figure 2
Two Spectrograms of the Utterance "Line up at the screen door." The upper spectrogram is the author's speech. The lower spectrogram is synthetic speech produced by DECtalk Software.

Replacing DECtalk

- Terry Sejnowski and Charles Rosenberg were interested in replacing DECtalk with a neural network
- Can one replace the thousands of rules and irregular examples of DECtalk with a small, distributed network that can handle regular and irregular pronunciations?



The New York Times — Technology

Learning, Then Talking

NETTALK, an artificial neural network that can be trained to pronounce English words, consists of about 300 neurons arranged in three layers - an input layer, which reads the words, an output layer, which generates speech sounds, or phonemes, and a middle, "hidden layer," which mediates between the other two.

The neurons, which are simulated on a computer, are joined to one another with 18,000 synapses, adjustable connections whose strengths can be turned up or down.

At first these volume controls are set at random and NetTalk is a structureless, homogenized tabula rasa. Provided with a list of words, it babbles incomprehensibly. But some of its guesses are better than others, and they are reinforced by adjusting the strengths of the synapses according to a set of learning rules.

After a half day of training, the pronunications become clearer and clearer until NetTalk can recognize some 1,000 words. In a week, it can learn 20,000.

NETtalk

- NETtalk was a neural network trained to generate the same I/O behavior as DECtalk
- 7 groups of 29 input units per group represent letters
- 80 hidden units
- Task: generate phoneme for middle input group
- Trained with generalized delta rule on corpus of 1024 words, informal text

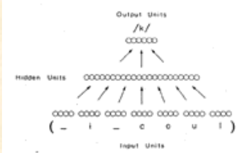


Figure 2 Schematic drawing of the network architecture. Input units are shown on the bottom of the pyramid, with 7 groups of 29 units in each group. Each hidden unit in the intermediate layer receives inputs from all of the input units on the bottom layer, and in turn sends its output to all 26 units in the output layer. An example of an input string of letters is shown below the inputs groups, and the correct output phoneme for the middle letter is shown above the output layer. For 80 hidden units, which were used for the corpus of continuous informal speech, there was a total of 309 units and 18,629 weights in the network, including a variable threshold for each unit.

NETtalk Learning

- NETtalk achieved 90% performance after being trained on only 5000 stimuli
- Klatt was not completely impressed: "In some sense, this is a surprisingly good result in that so much knowledge could be embedded in a moderate number of about 25,000 weights, but the performance is not nearly as accurate as that of a good set of letter-to-sound rules" (Klatt, 1987, p. 770)

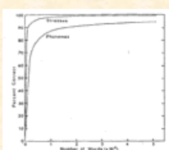


Figure 3 Learning curves for phonemes and stresses during training on the 1024 word corpus of continuous informal speech. The percent of correct best guesses are shown as functions of the number of training words.

NETtalk Representations

- Because of its size and complexity, the internal structure of NETtalk was not investigated in detail
- Sejnowski and Rosenberg explored the hidden layer with Hinton diagrams
- They concluded that most of the representations were 'distributed'

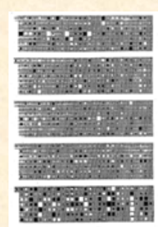


Figure 5 Hinton diagram showing weights from the layer of input units to 5 representative hidden units taken from a network with 80 hidden units that was trained on a corpus of continuous informal speech. Each gray rectangle represents one hidden unit and each square within a rectangle represents a weight. The area of the square is proportional to the magnitude of the weight and the sign of the weight is indicated by the color: white for positive, or excitatory weights, and black for negative, or inhibitory weights (38). The largest weights have magnitudes of about 4. Each row of squares within a gray rectangle represents the weights from one group of input units, with the leftmost input group on the top row and the rightmost input group on the bottom row. The isolated weight in the upper left corner of each array is the bias (negative threshold) of the unit, which was treated like a weight to a true unit. Out of 29 squares in each row, the first 26 represent the weights from the letters of the alphabet, from a to z, and the last three represent the punctuations, including word boundaries. Thus, the square in the lower left corner of a hidden unit is the weight it receives when the letter "a" is present in the rightmost input group. For most hidden units more than one combination of letters will cause it to produce a large output. These are called distributed representations. However, a few hidden units, such as the third from the top, had a more restricted pattern of weights that could be called a local representation.

NETtalk Develops

- "During the early stages of learning in NETtalk, the sounds produced by the network are uncannily similar to early speech sounds of children"
  - [Examples of NETtalk]
- "The phonological mappings produced by NETtalk are efficient encodings for a parallel network and may be comparable to those used by humans"
- Descendants of NETtalk have been central in the debate about the kinds of model required to account for reading, as well as symptoms of dyslexia

Terry Sejnowski

Dyslexia

- Dyslexia is a disorder in reading of words, and can be related to brain injury
  - Phonological dyslexia is a disorder in which nonwords cannot be read, but the reading of words is unaffected
  - Surface dyslexia is a selective disorder in which there is severe difficulty in reading aloud irregular words, usually revealed in terms of generalization errors; nonwords can be read
  - Deep dyslexia involves semantic errors in reading aloud, visual errors, and an inability to read nonwords

Symptoms of Dyslexia

- Intelligent but has problems with reading, writing or spelling
- Often confuses the right from the left
- Intelligent but does not test well or has severe anxiety about testing
- Seems to daydream or zone out when in a classroom or meeting scenario
- Learns best by "hands on" training rather than verbal or written instruction
- Sees movement of letters on a page whether reading or writing
- Reads and rereads without much comprehension
- Has difficulty with spelling
- Has challenges putting thoughts into words
- Difficulty with writing or copying
- Tends to hold a pen or pencil differently and very tightly
- Handwriting is hard to read
- Has difficulty with large or fine motor skills
- Has difficulty reading time on a traditional clock
- Has time management problems
- Tends to be a procrastinator
- Tends to be good at math calculations but word problems are very difficult
- Tends to be disorderly or extremely orderly

EDUCATE

Deep Dyslexia

Deep dyslexia's symptoms are difficult to explain using simple boxologies

1. Semantic errors (e.g., BLOWING "wind", VIEW "scene", NIGHT "sleep", GONE "lost");
2. Visual errors (e.g., WHILE "white", SCANDAL "sandals", POLITE "politics", BADGE "bandage");
3. Function-word substitutions (e.g., WAS "and", ME "my", OFF "from", THEY "the");
4. Derivational errors (e.g., CLASSIFY "class", FACT "facts", MARRIAGE "married", BUY "bought");
5. Non-lexical derivation of phonology from print is impossible (e.g., pronouncing nonwords, judging if two nonwords rhyme);
6. Lexical derivation of phonology from print is impaired (e.g., judging if two words rhyme);
7. Words with low imageability/concreteness (e.g., JUSTICE) are harder to read than words with high imageability/concreteness (e.g., TABLE);
8. Verbs are harder than adjectives which are harder than nouns in reading aloud;
9. Functions words are more difficult than content words in reading aloud;
10. Writing is impaired (spontaneous or to dictation);
11. Auditory-verbal short-term memory is impaired;
12. Whether a word can be read at all depends on its sentence context (e.g., FLY as a noun is easier than FLY as a verb).

Dual Route Cascade Model

- Coltheart's dual route cascade model (DRC) is a classical model of reading
- Basic assumption: there are multiple routes by which text can be converted into speech, some involve semantics, others do not
- Damage to different routes in this model can account for different kinds of dyslexia, and can account for the un-unified syndrome of symptoms associated with deep dyslexia

Max Coltheart

Evolving From NETtalk

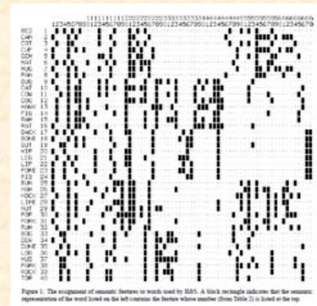
- The success of NETtalk paved the way for other researchers to explore networks that converted text into something else
- Geoffrey Hinton and Tim Shallice, for instance, began to study networks that were models of reading
- These networks mapped, for example, graphemes to phonemes – but included intermediate semantic representations too
- Issue was whether such models could provide an alternative to classical, multiple route models, like Coltheart's DRC

Geoffrey Hinton

Tim Shallice

Primitives

- Hinton and Shallice began with a small set of primitive features – letters, words, and semantic features, and defined mappings between them



A Variety Of Architectures

- Hinton and Shallice explored a number of different network architectures to map one kind of feature into another, all motivated as models of reading
- Key architecture mapped graphemes through semantics to phonemes

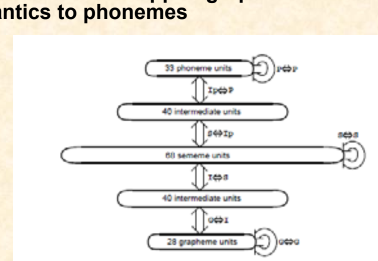
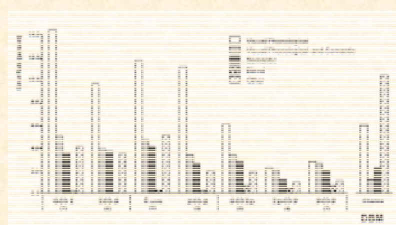


Figure 12. The DBM architecture for mapping among orthography, semantics, and phonology.

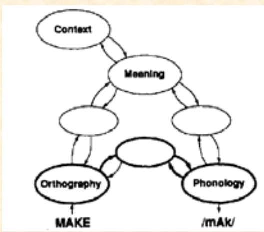
Connectionist Neuroscience

- Hinton and Shallice explored the effects of a variety of lesions of their networks after training was completed
- They produced errors associated with deep dyslexia



Single Route Models

- Surface vs. deep dyslexia have led to dual route models, similar in structure to DECtalk
- Plaut, Seidenberg, Shallice and others suggest connectionist models provide single route theories that can account for various types of dyslexia



David Plaut

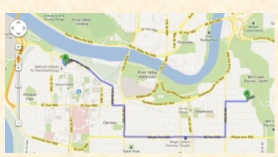
Mark Seidenberg

Case Study 2: Metric Properties Of Space

- Define some distance function d(x,y) that delivers the distance between points x and y
- The set of points that can be input to this function is a metric space if:
  - $d(x,y) \geq 0 = d(x,x)$ (minimality)
  - $d(x,y) = d(y,x)$ (symmetry)
  - $d(x,y) \leq d(x,z) + d(z,y)$ (triangle inequality)

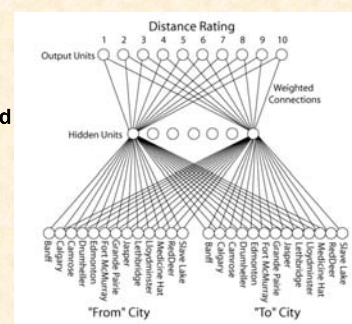
Is Physical Space Metric?

- In terms of traditional distance measures, physical space is metric
- However, alternative measures of distance make physical space nonmetric
  - If distance = time traveled, then physical space is nonmetric because it violates the symmetry constraint
  - If I drive, I can get home faster than I can get to work



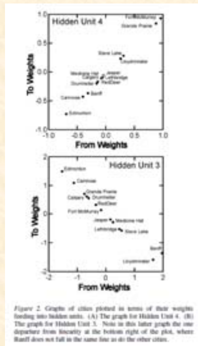
Networks And Metric Space

- Can artificial neural networks represent metric properties of space?
- Dawson, Boechler & Orsten (2005) studied this issue with a network of value units
- The network was trained to make judgments of distances between Albertan cities
- The judgments obeyed the three metric principles



A Converged Network

- With 169 training patterns, the network converged in 5078 sweeps
- It had internalized a metric space
- The weights at the end of training were highly systematic
- However, the weights did not appear to represent distances!
- For instance, weights could not be used to predict distances on the map



Hidden Units In Space

- Dawson et al. (2005) were able to treat hidden units as if they occupied particular locations on the map of Alberta
- Key property: Hidden units defined a plane with a particular direction of view
- <u>Direction</u> was how distance was being measured!
- "Imagine being on a flat prairie with a small number of distinct landmarks (e.g., trees) in view, and having the task of determining the distances between all of these landmarks. However, the only tool available was a sextant, so that the only measurement that could be taken was the angular displacement between pairs of landmarks. One could make a rough estimate of the distances between landmarks by standing at one location on the prairie and taking a sextant reading between every possible pair of landmarks. The reasoning would be that if a sextant reading was high, then the two landmarks were far apart, and that if it was near zero, then the two landmarks were near one another" (Dawson et al., 2005, p. 44)

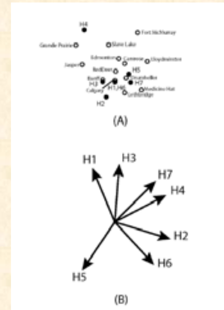
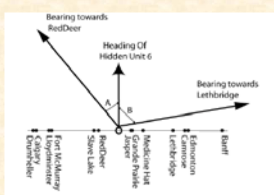


Figure 3. (A) Locations of hidden units relative to cities on the map of Alberta. Hidden units are indicated with solid dots. (B) Headings of each hidden unit, centered to a common origin.

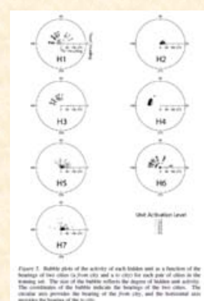
Hidden Unit As Sextant

- Each hidden unit could be seen as a sextant, delivering angles or bearings towards pairs of cities
- Connection weights were strongly correlated with this model
- But this means that each hidden unit delivers an inaccurate distance measure



Coarse Coding

- Individual inaccuracy is dealt with by having multiple views (multiple hidden units with different bearings)
- Pooling these inaccurate, but varied, responses together generates accurate distance readings
- A committee of sextants!
- This kind of coding should be able to cope with violations of metric properties!



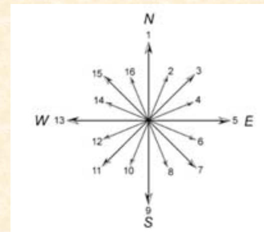
Case Study 3: Nonmetric Space

- **Judgments of similarity are not symmetric**
  - The judged similarity of North Korea to China exceeds the judged similarity of China to North Korea
- **Judgements of similarity violate the triangle inequality**
  - Jamaica is similar to Cuba
  - Cuba is similar to Russia
  - but Jamaica is not similar to Russia at all!

Amos Tversky

Antisymmetric Space

- Direction is a spatial relation that is a radical violation of symmetry – it is antisymmetric
- This is because if *d(x,y)* delivers direction, then *d(x,y) = - d(y,x)*
- Dawson & Boechler (2007) explored a multilayered network of value units that was trained to deliver directional judgments
- Given two cities, deliver the compass rose direction from one to the other



Asymmetric Training

- Again, a network with 7 hidden units, trained on 169 patterns, converged after 7645 sweeps of training
- Hidden unit behavior reflected the asymmetry of the task
- Hidden units in a 13 x 13 city matrix had large asymmetries of both net inputs and of activities

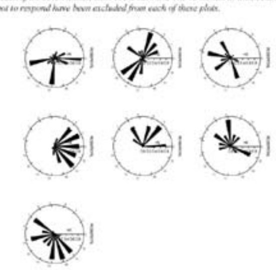
Table 2. Measures of asymmetries of the activations and net inputs of the hidden units to each pair of cities, and the correlation between the two sets of weights that feed into each hidden unit. See text for details.

Hidden Unit	Proportion Asymmetry Of Activation Matrix	Proportion Asymmetry Of Net Input Matrix	Correlation Between "From" Weights and "To" Weights
H1	0.47	0.63	-0.27
H2	0.36	0.36	0.28
H3	0.51	0.49	0.03
H4	0.92	0.95	-0.91
H5	0.72	0.86	-0.76
H6	0.45	0.50	-0.01
H7	0.81	0.92	-0.86

More Coarse Coding

- When hidden unit activity was plotted in terms of two cities in the context of a preferred direction, it was clear that the hidden unit system coarse coded output direction

Figure 2. Bar plots of the median hidden unit activity as a function of the petal on the compass rose, for each of the seven hidden units in the network. The 13 stimuli to which the network was trained not to respond have been excluded from each of these plots.



Coarse Coding Again

- Coarse coding is further revealed by looking at patterns of activity when city pairs are presented that all map onto the same output direction
- Interaction of three hidden units is required; simple local feature detection is not evident!

Table 3. Activation produced in three different hidden units by the 15 city pairs that all cause the network to choose compass petal 6 as the response

"From" City	"To" City	H3	H5	H6
BANFF	LETHBRIDGE	0.96	0.41	0.01
BANFF	MEDICINE HAT	0.96	0.46	0.01
CALGARY	MEDICINE HAT	0.06	0.87	0.61
GRANDE PRAIRIE	CAMROSE	0.17	0.77	0.85
GRANDE PRAIRIE	DRUMHELLER	0.64	0.75	0.60
GRANDE PRAIRIE	EDMONTON	0.89	0.99	0.97
GRANDE PRAIRIE	LLOYDMINSTER	0.01	0.49	0.39
GRANDE PRAIRIE	MEDICINE HAT	0.66	0.72	0.23
GRANDE PRAIRIE	RED DEER	0.00	0.75	0.81
JASPER	CALGARY	0.72	0.86	0.49
JASPER	DRUMHELLER	0.37	0.57	0.41
JASPER	LETHBRIDGE	0.68	0.66	0.37
JASPER	MEDICINE HAT	0.35	0.61	0.38
RED DEER	DRUMHELLER	0.78	0.56	0.00
SLAVE LAKE	LLOYDMINSTER	0.61	0.76	0.53

Head Direction Cells

- The hidden units in the network are analogous to head direction cells
- These cells are also coarsely tuned
- Current theories combine these cells into a system of overlapping sensitivities, as coarse coding would require
- The bottom figure shows networks of cells; the greyer the cell the higher the activity.
- Head direction is mediated by parallel processing