Introduction to speech perception

Linguistics Paper 9
Foundations of Speech Communication
Sarah Hawkins
5: 7 November 2007
Aim and claim

• To introduce some of the “facts” of speech perception in the context of pattern perception

• Starting with the claim that all perceptual activity involves brain processes that
  – mesh together sensation with memory
  – via prediction of likelihood (probability) that current sensation matches memoryA better than memoryB.
• “To predict the next note of a song, you need to know the song’s name, where you are in the song, how much time has passed since the last note, and what that last note was.” (Jeff Hawkins, 2004:141).

• Implies that perceptual processing of stimulus A depends heavily on experience and its perceived context as well as its absolute properties: relationships are paramount (Sarah Hawkins, 2004)
Apparently random relationships between properties of the signal
...can become meaningful in the right context

but only with the right experience

and when it meshes with circumstances and expectations

Bregman, 1990
The brain as pattern processor: 1

- **Localization of function:** Areas appear to specialize in a particular modality, or a particular type of processing.

- Sensory information has been mainly seen as going from the periphery of the body to successively "higher" brain centres.

Colours = illustration of areas roughly associated with some basic functions. Numbers are Brodmann’s areas, which are defined in terms of cell structure, not function.
Areas appear to specialize in a particular modality, or a particular type of processing, but
- there is much “plasticity”: e.g. sign language activates the primary auditory cortex
- all interact with one another

Sensory information is traditionally seen as going from the periphery to successively “higher” parts of the brain, but there are many more backward connections than forward connections
- poorly understood, but of great interest
A final claim

Nothing is interpreted in isolation from experience and expectation

recognition occurs when memory and sensation match
Speech perception as pattern perception

stringing together simple percepts, or a complex mental construction from whatever information is available?
Sine wave speech

When sine waves are combined to mimic the centre formant frequencies and temporal patterns of natural speech, most people hear …… ??

Robert Remez et al.
Sine wave speech

1. I read a book today
2. These days, a chicken leg is a rare dish
3. The boy was there when the sun rose
4. Where were you a year ago?
5. Where were you a year ago?  F2 only

Robert Remez et al.
Sine wave speech

• Is this to be expected, the result of reproducing essential but simple patterns, or is there more to it?

• Listeners usually cannot choose whether to hear sounds as speech or as nonspeech. But they can be primed, and some stimuli will be heard as speech by some individuals and as nonspeech by others.

• Once you have heard a signal as speech, it is virtually impossible to hear it as nonspeech. “Modularity”?

• (Though some specially-designed stimuli can be heard as both speech and nonspeech simultaneously.)
Speech perception as context-dependent

expectations

formed from past experience
Vowel-to-vowel coarticulation

\(/ib\text{\v{e}b}i/ vs /ab\text{\v{e}b}\text{\v{a}}b/\)

Naturally spoken

Schwas exchanged

\(/ib\text{\v{e}b}i/ /ab\text{\v{e}b}\text{\v{a}}b/\)

\textit{this experience: language-specific, phonological-phonetic}

\textit{but no meaningful context}
Ladefoged and Broadbent (1957)

"Please say what this word is:
  bit  bet  bat  but

Formant frequencies of the target word are identical in the two contexts. Only the formant frequencies of the carrier sentence vary.

this experience: speaker-specific (phonological, phonetic, grammatical, meaningful); long-domain...

F1 of CARRIER

bet  200-380 Hz

bit  380-660 Hz
Other types of knowledge that drive expectations

experience tells you what is possible and what is likely

efficient brain processing lets you access memories of those experiences
The context of possible responses affects intelligibility

- monosyllables
- size of test vocabulary affects identification
  - 2...256...all monosylls
- though presumably there are limits:
  - two vs six
  - five vs nine!

Miller, Heise & Lichten, (1951) *J.Exp.Psych. 41*, 329-335
Multi-modal speech perception:

Sumby & Pollack (1954) (JASA)

Especially in high levels of noise:

• audiovisual presentation increases intelligibility (visual contribution is relative to the available auditory contribution)

• (much subsequent work confirms this; it is currently a hot topic in speech technology)
Multi-modal speech perception: the McGurk effect

- the effect is not just a case of enhancing or supporting the auditory signal with another sensory signal
- it reflects integration of information across sensory modalities (here, vision and audition)
- the perceptual response reflects knowledge of what is possible and likely

McGurk and MacDonald (1976) *Nature*, 264, 746-748
Multimodal speech perception:
the McGurk effect

Listeners heard CVCV nonsense syllables and
saw video of lips saying different syllables (time-synchronised)

<table>
<thead>
<tr>
<th>STIMULUS</th>
<th>RESPONSE (adults)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hear</td>
<td>See</td>
</tr>
<tr>
<td>(a)</td>
<td>baba gaga</td>
</tr>
<tr>
<td>(b)</td>
<td>gaga baba</td>
</tr>
</tbody>
</table>

In (b), both heard and seen information is distinctive; but in (a), both heard and seen information is ambiguous. The resultant percepts are quite different.

Data from McGurk and MacDonald (1976) *Nature*, 264, 746-748
The McGurk effect

- The McGurk effect is one of many demonstrations that *all salient information* seems to be used to understand speech, and presumably meaning.

- Responses seem to reflect perceived reliability of information from many potential influences: each type of information contributes in proportion to its perceived quality/reliability (to be discussed in later sessions).
What is likely: Semantic predictability

Relation between intelligibility and signal-to-babble ratios for two different degrees of predictability determined from preceding context: the SPIN test (used to test hearing in the elderly).

Predictability low, e.g. *We were talking about the road*

Predictability high, e.g. *The car drove down the road*
Semantic predictability: the SPIN test

+10 dB S/B
- She had spoken about the scar
- Camels store water in their humps
- Mr. Smith might discuss the mill
- I built the model from a kit
- The loud noise made him jump with fright

Kalikow, Stevens, and Elliott J. Acoust. Soc. America 1977
Semantic predictability: the SPIN test

0 dB S/B Form 2.8
- The bird of peace is the dove
- Tom had spoken about the pill
- The cigarette smoke filled his lungs
- They’ve considered the sheep
- Cut the meat into small chunks

Kalikow, Stevens, and Elliott J. Acoust. Soc. America 1977
Semantic predictability: the SPIN test

-5 dB S/B    Form 2.1

- The watchdog gave a warning growl
- She made the bed with clean sheets
- The old man discussed the dive
- Bob heard Paul called about the strips
- I should have considered the map

including knowing what to listen for

Kalikow, Stevens, and Elliott J. Acoust. Soc. America 1977
Some other linguistic types of predictability/knowledge

- **phonotactics** (permissable/probable sound sequences)
  - e.g. syllable-initial / spr spl str skr (skl) / but not / stl /

- **word frequency** Savin (1963) high/low frequency words heard in white noise. Higher freq words recognised at lower signal-to-noise ratios than low freq words

- **neighborhood density** (usually measured, clumsily, as number of words with n phonemes different from the current word (n is usually 1)
  - e.g. square, cigarette (low ND) vs bat, rugger (high ND)

- **grammatical markers, grammatical agreement**
Patterns in Polysyllables constrain the word choices

- polysyllables (cigarette, cathedral) are more intelligible than monosyllables (in auditory-only presentations)
  Sumby & Pollack (1954) JASA 26: 212-215

distinctive overall shape...
  neighborhoods...
  cohorts...
Patterns in Polysyllables constrain the word choices

• neighborhood density (ND) counted in phonemes rubber rudder rugger runner dubber.... affects word recognition
  – words from sparse neighbourhoods more quickly and accurately identified than those from dense neighbourhoods

• but ND is a crude measure: does not apply as well to polysyllables or consonant clusters
  – it is worthwhile speculating as to why, including for connected speech
Patterns in Monosyllables: “Holes” in VC# patterns

<table>
<thead>
<tr>
<th>Common</th>
<th>Rare</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>/Vd/</td>
<td>/ab/</td>
<td>/ag/</td>
</tr>
<tr>
<td>hard</td>
<td>harp</td>
<td>card</td>
</tr>
<tr>
<td>card</td>
<td>carp</td>
<td>/ab/</td>
</tr>
<tr>
<td>heed</td>
<td>heap</td>
<td>heat</td>
</tr>
<tr>
<td>seed</td>
<td>seep</td>
<td>/ig/</td>
</tr>
<tr>
<td>weed</td>
<td>weep</td>
<td>/ib/</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Do such “holes” make perception easier? What process would underlie that?
- listeners ‘dismiss’ such sequences as unlikely?
- listeners fail to identify them because they have not been heard before?
Summary

• This lecture has demonstrated some classical effects of speech perception, each aimed at illustrating how sensation and memory (experience, knowledge) combine to form a particular percept.

• Later lectures will discuss these and other effects in terms of
  – perceptual formation of linguistic-phonetic categories
  – the sorts of processes needed to model this: paradigmatic, syntagmatic, hierarchical,....?
  – properties of speech that suggest this type of analysis may be valid

• Demonstrations of the power of experience on percept:
  http://www.richardgregory.org/experiments/index.htm
Reading

1. For introductions to speech perception
   Reading list section 6.1 and 6.2
   Note Hawkins (2004) paper is downloadable in two forms:
   - the official published version:
     http://www.rle.mit.edu/soundtosense/conference/pages/invited.htm
   - and the unofficial version with extra figures:
     http://www.ling.cam.ac.uk/sarah/docs/Hawkins-SoundToSense_submission2.pdf

2. Introduction to brain functioning:
   Easy reading, good ideas. The best accessible resource that I know of for
   the prediction/memory/pattern perception approach
Advanced reading

   Phonological features, auditory objects, and illusions.
   Ask me for a pre-publication copy if interested.

Categorical Perception
Stop consonants: place of articulation

Equal acoustic changes → unequal auditory percepts

place of articulation of stops: /b/ vs /d/ vs /g/

Liberman, Harris, Hoffman, and Griffith (1957)
Journal of Experimental Psychology 54, 358-368
Categorical perception of stop place

Boundaries in identification function coincide with peaks in discrimination function

Left: the percentage of time each stimulus was identified as /b/, /d/ or /g/.
Right: two-step discrimination test.
Categorical perception of stop consonants: VOT

Waveforms of three stimuli from a continuum where the only thing varied was the duration between the burst and the onset of periodic excitation (VOT), which increased in 10 ms steps from -40 ms to +100 ms.

Stimuli used in first demo of CP in newborn infants (4 weeks old).