Introduction to speech perception and linguistic categories

Paper 9
Foundations of Speech Communication
Michaelmas Week 3

Katharine Barden
Aim

• To consider how we seem to recognise discrete linguistic units from the speech signal.

(Note the assumption that mental linguistic units are discrete.)

• To show how categories can emerge from a combination of “bottom-up” and “top-down” processing
3 ways discrete linguistic-phonetic categories might be perceived

1. they are really there in the acoustic signal
   acoustic invariance (or reliability)

2. they result from the way the auditory system processes sound
   auditory invariance (or reliability)

3. they result from the way the brain processes any sort of information, sensory or not
   ‘cognitive invariance’
Classical assumptions

Classical theories held/hold that:

• We necessarily perceive **phonemes** or **phonological distinctive features** when we listen to speech…

• …because each phoneme or feature bears an **invariant relationship** with some property/properties of the speech signal
Do linguistic units have invariant acoustic correlates?

Sources of variation we have come across so far...

- syllable position e.g. ‘clear’ and ‘dark’ /l/
- register e.g. careful/casual
- grammar e.g. function/content
- position in interaction e.g. turn-taking
- talker e.g. voice quality, male/female, pathological speech
Do linguistic units have invariant acoustic correlates?

- There is no strong evidence that all linguistic units, even of a single type, are invariantly and reliably present in the speech signal.

- Yet some acoustic-phonetic features are more robust across contexts and speakers than others.
Robust features: spectrogram of “My family lives in Oxford”
Robust features (e.g. Zue, 1985)

- "Strong" fricative — "weak" fricative — nasal — periodic — silence — transient — vowel (high/low, front/back, spread/round)

- These offer a set of "invariant" acoustic features from which to make preliminary decisions about what words were spoken

- Some Automatic Speech Recognition (ASR) techniques use such broad featural categories; less widely applied to human speech perception work
Robust features

- Usually clearly visible in spectrograms
- Independent (e.g. you can know it’s a “strong fricative” without knowing its exact place of articulation)
- BUT only work for a small set of features that have simple acoustic properties. They don’t tell us place or voicing of stops, for example.
Relational invariants

• Defining categories in relational terms can be useful than looking for absolute acoustic correlates:
  – Some relations are static e.g. vowels with two formants close together in frequency
  – Some are dynamic e.g. how phonetic characteristics relate to each other over time
Dynamic relational invariants for stop place of articulation? (Stevens)

- Bilabial
  - burst flat or falling, low amp
- Alveolar
  - rising: burst > vowel spectrum at high freqs
- Velar
  - compact mid-freq peak near F2 & F3

Onset of vowel [ɛ]
Summary: Relationships between properties of the signal are critical

• Current views are that relationships between successive acoustic (and visual) events define linguistic categories as much or more than static properties

• Listeners interpret sensory information (e.g. acoustic and visual input) in terms of relationships between properties that reflect the coordinated, dynamic behaviour of the vocal tract
3 ways discrete linguistic-phonetic categories might be perceived

1. they are really there in the acoustic signal
   acoustic invariance (or reliability)

2. they result from the way the auditory system processes sound
   auditory invariance (or reliability)

3. they result from the way the brain processes any sort of information, sensory or not
   ‘cognitive invariance’
What is the relationship between the physical parameters of a sound and the percept?

Can auditory processing and low-level perceptual processing produce invariant/reliable correlates of linguistic units?
What is the relationship between the physical parameters of a sound and the percept?

- Not a simple linear relationship between physical parameters and how we perceive them.
  - e.g. fundamental frequency and pitch
    An octave represents a doubling of fundamental frequency, yet octaves are heard as the same size pitch jump no matter what absolute frequency changes are involved.
Weber’s law

• Our ability to detect changes in a parameter depends on the magnitude of the reference

• The smallest detectable change in a stimulus is proportional to the magnitude of that stimulus
Audible range of frequencies

- Tuna: 50-1,100 Hz
- Chicken: 125-2,000
- Goldfish: 20-3,000
- Human: 64-23,000
- Dog: 67-45,000
- Cat: 45-64,000
- Mouse: 1,000-91,000
- Bat: 2,000-110,000
- Beluga whale: 1,000-123,000
- Porpoise: 75-150,000
Loudness depends on frequency

Threshold of Audibility of Pure Tones

Source: Mark Huckvale
UCL web site
Tutorials 2005
Islands of auditory reliability?

Syrdal and Gopal (1986, JASA 79, 1086-1100):

Left panel: Scatterplot on a linear frequency scale of F1 frequency versus F2 frequency for American English vowels spoken by men, women, and children (data from Peterson & Barney 1952).

Right panel: the same data, replotted on a Bark scale in terms of F3-F2 Bark frequency versus F1-f0 Bark frequency.
Islands of auditory reliability?

Some sounds are distinguished, and others are grouped together, because of the way the auditory system responds to them e.g. dimensions of vowel quality
3 ways discrete linguistic-phonetic categories might be perceived

1. *they are really there in the acoustic signal*
   
   **acoustic invariance (or reliability)**

2. *they result from the way the auditory system processes sound*
   
   **auditory invariance (or reliability)**

3. *they result from the way the brain processes any sort of information, sensory or not*

   ‘**cognitive invariance**’
Categorical Perception
Stop consonants: place of articulation

Equal acoustic changes → unequal auditory percepts

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).
How ‘categorical’ is Categorical Perception?

- Category boundaries are not stable, but highly *labile*: they shift under the influence of many different factors.
CP boundary shifts: range effects

- identification expt e.g.
- VOT continuum da.........ta
- when stimuli are removed from one end, the 50% id boundary shifts towards the other

![Graph showing boundary shifts](image-url)
Rate effects on CP boundary

- identification expt e.g.
- VOT continuum
da..........ta

- when stimuli are faster, the 50% id boundary shifts

CP boundary shifts: cue trading

- Cue trading: more of one property compensates for less of another

  e.g. for stimuli whose VOT is ambiguous between /da/ and /ta/, decreasing burst amplitude causes more /da/s to be perceived

Cue trading is also called trading relations
Lexicality (meaning): “Ganong effect” on CP boundary

- Identification experiments (VOT series)
- Word vs nonword at ends
- More /d/ responses if real word begins with /d/ (dash—tash)

Perception is more forgiving when the sound means something!

Effect of sentence meaning on CP boundary

• Similar boundary shifts for sentence meaning…

  e.g. *The farmer milked the [g/k]oat in the dairy*

Summary: CP boundary effects

Perception adjusts to the distribution of stimuli & is more forgiving about unclear sounds if the message makes sense.
When expectation overrides sensation

Charlie Chaplin:
http://www.richardgregory.org/experiments/index.htm
Sine wave speech

Sine waves are combined to mimic the centre formant frequencies and temporal patterns of natural speech.

Robert Remez et al.
Sine wave speech

• Listeners usually cannot choose whether to hear sounds as speech or as nonspeech, but they can be primed.

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

• Role of expectations:
  – priming – speech vs. non-speech
  – prior knowledge of what will be said makes it easier to interpret the signal
Other types of knowledge that drive expectations
The context of possible responses affects intelligibility

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

Miller, Heise & Lichten, (1951) *J.Exp.Psych.* 41, 329-335
Semantic predictability: the SPIN test

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).

Klikow, Stevens, and Elliott J. Acoust. Soc. America 1977
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

Kalikow, Stevens, and Elliott J. Acoust. Soc. America 1977
Semantic context: phonemic restoration

• Restoration of different phonemes depending on context
  – *eel was on the orange
  – *eel was on the axle
  – *eel was on the table

• Demonstrates effect of semantic context, and “backwards” processing

Warren & Warren 1970
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. *bat* (high ND) vs *cigarette* (low ND)
  - words from sparse neighbourhoods more quickly and accurately identified than those from dense neighbourhoods (Luce and Pisoni, 1998) Ear and Hearing, 19, 1-36)
Influence of the surrounding phonetic context on perception
Vowel-to-vowel coarticulation

/ibəbi/ vs /abəba/

Naturally spoken

Schwas exchanged

/ibəbi/ /abəba/
“BARB a barb” vs “BEEB a beeb”

/babəbab/  /bibəbib/
"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.

Ladefoged and Broadbent (1957) JASA 29, 98-104
Multi-modal speech perception: the McGurk effect

- Demo at:

http://www.youtube.com/watch?v=aFPtc8BVdJk

Multi-modal speech perception: the McGurk effect

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

• Demonstrates integration of information across sensory modalities (here, vision and audition)

Multimodal speech perception: the McGurk effect

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

In (a), both heard and seen information is ambiguous; but in (b), both heard and seen information is distinctive. 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.

• The perceptual response reflects knowledge of what is possible and likely.
3 ways discrete linguistic-phonetic categories might be perceived

1. they are really there in the acoustic signal
   
   \textit{acoustic invariance (or reliability)}

2. they result from the way the auditory system processes sound
   
   \textit{auditory invariance (or reliability)}

3. they result from the way the brain processes any sort of information, sensory or not
   
   \textit{‘cognitive invariance’}
All three?

Some acoustic features have properties that are a good deal more invariant and reliable than others, especially if they are considered with respect to known properties of the auditory system, and in relation to their surrounding context, whether that context is phonetic, visual or relates to linguistic knowledge and expectations.
Summary

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

• The process of speech perception: stringing together simple percepts, or a complex mental construction from whatever information is available?
Next term

- How are phonetic categories acquired?
- Plasticity of phonetic categories
- Consider whether ‘knowledge’ may be formed from statistical distributions, built up from repeated experience of the way the information occurs in the speech signal
- How are categories represented in the brain?
Reading

For introductions to speech perception

Reading list section 6.1, 6.2 and 6.3

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

For categorical perception, see reading list section 6.4
Advanced reading

Hawkins, S. Phonological features, auditory objects, and illusions. 
*Journal of Phonetics* (in press)
