Understanding Speech 9C

1. Speech perception – how we identify sounds of language
   1. Think about how much information goes into understanding someone speaking
   2. (swamp people clip)
   3. Information from cognitive notes
2. Recognizing speech
   1. Why is speech perception difficult?
      1. Visual word perception – you have all day to recognize the words in front of you, spoken word perception – it’s said and then is gone.
      2. Sounds are harder to segment, because it’s not like we pause between words and sounds slur together.
   2. We are fairly good at this task though, processing is automatic
      1. Can understand at about 20 phonemes a second or faster
      2. Can recognize words after about 200 msec
      3. Context helps us figure it out
      4. We can select words over background noise
   3. Acoustic signals and phonetic segments: how do we segment speech?
      1. Acoustic sounds are not fixed – they depend on context (b = ball bill able rob), so understanding them is complex
      2. Templates – we don’t have a perfect example for recognizing sounds – kind of like how there are many ways to write letters (template or feature matching theory)
      3. Features
         1. Segmentation – splitting speech up into constituent phonemes
         2. Invariance – same phoneme can sound different depending on context
      4. Invariance
         1. Assimilation – influence of one sound on the articulation of another, so the sounds blend together
         2. Co-articulation – the way in which the articulatory apparatus takes account of the surrounding sounds when a sound is articulated; sound conveys information about its neighbors (two sounds at once)
         3. Actually has advantages – each sound is easier to say as we blend, and we are getting more information at once (several phonemes at a time)
            1. Basically tells us what’s coming next – parallel transmission
      5. Segmentation
         1. Words run together as well as sounds, so there’s a lot of segmentation going on (think about a movie in a foreign language and how it all runs together)
         2. Possible – word constraint – we break up what we are hearing into possible words, don’t break up in the middle of words leaving syllables hanging
         3. Syllables – metrical segmentation strategy – using this information
            1. Strong syllables have stress and tend to indicate content words
            2. Weak syllables tend to be function words
            3. (conducts ascents uphill – how do you hear it?)
         4. Stress based segmentation – basing word boundaries as stressed syllables
         5. Syllable based segmentation – looking for the syllables, better in other languages
   4. Categorical perception
      1. Definition – perceiving things that lie along a continuum as belonging to one distinct category or another
      2. We classify phonemes based on this strategy even though that’s not the way they work
         1. We tend to hear b p d as different even though you might be hearing a blend of the two (mcgurk?)
      3. Voice onset time – how long it takes for the vocal chords to start vibrating (b d voiced) (p t delayed)
      4. We probably understand stuff on the continuum and are sensitive to it, but based on the system we are using, we force things to be categorical
   5. Prelexical code – do we need to identify phonemes before the words?
      1. Some research shows that we are better at responding to syllables than individual phonemes and that we only recognize phonemes later
      2. Dual code theory
         1. Prelexical – phonetic code
            1. Computed from the perceptual analysis
         2. Postlexical – phonemic code
            1. Computed from the top down word analysis
      3. Phoneme monitoring task – press a button as soon as you hear a particular sound
         1. Words and nonwords are the same – so we are responding to prelexical code
         2. But if you include context – the postlexical code interacts (he sat reading a book/bill until it was time to go home – faster at book because it’s expected)
      4. Summary – still debate about whether we need to identify phonemes first.
3. Role of context in sound identification
   1. Speech perception is interactive – both top down and bottom up processes influence your perception of the words spoken.
   2. Exp 1
      1. Varied the sounds between k and g, but stuck them on the –iss rime.
      2. People would only say it was kiss because giss is not a word
      3. Lexical identification shift – words influencing our categorical perception of an ambiguous phoneme
   3. Signal detection exp – (explain signal detection theory)
      1. So signal detection allows you to see the point at which something becomes different (or apparent), when you notice the change
      2. Influenced by both the word context (i.e. above) and the sentence context
   4. Phonemic restoration exp
      1. Participants were given sentences with a cough or missing segment in a sentence.
      2. They would report they heard the phoneme even if they knew they didn’t.
      3. Word reports depending on the surrounding context of the sentence.
         1. The \*eel was on the orange.
         2. The \*eel was on the axle.
         3. The \*eel was on the shoe.
         4. The \*eel was on the table.
      4. Some other research that says that people guess at what phoneme is missing, but then use postlexical information to figure out which phoneme makes the most sense.
      5. So sentence context isn’t affecting that prelexical processing, just helping you figure out the meaning and biasing you towards which phoneme is right.
4. Time course in spoken word recognition
   1. Identification stages
      1. Initial contact
         1. Some representation of the sensory input makes initial contact with the lexicon
         2. Once the lexical entries begin to match, they get activated (think about connectionist models here)
      2. Lexical selection
         1. Sensory input continues to accumulate until one lexical entry is selected
      3. Word recognition
         1. Word is recognized and the recognition point usually occurs before the complete word has been heard.
         2. This is the uniqueness point – where it denotes that it’s one word and no other.
         3. Isolation point – when you’ve got the word but are waiting for the uniqueness point, basically context has biased you to this answer.
         4. Recognition point – confidence of the word answer
      4. Lexical access – point after recognition in which all information about a word becomes activated and available.
      5. Integration – putting all that information together and forming a representation of structure/syntax
   2. Frequency
      1. Seems to be important at very early stages of processing. People were shown pictures of high frequency and low frequency objects. (Bed bench bell). As soon as the b sound occurred they started looking at the bed because it’s higher in frequency. Only after another phoneme or two did they look at the right lower frequency object.
5. Context effects in word recognition
   1. Models for role of context
      1. Autonomous – context cannot have an effect until after word recognition
      2. Interactionist – information interactions, the effect of lexical information flows back down to sensory information
   2. Context
      1. Non-structural context – information from the same level of processing as what is currently being processed
         1. Kind of like priming information doctor activates nurse
      2. Structural context – affects the combination of words into higher units (top down processing)
         1. Semantic context based on word meanings
         2. Interpretative – high level information such as pragmatic information, discourse, knowledge about the world.
6. Models of speech perception
   1. Models
      1. Template matching – targets are stored as templates and recognition occurs when a match is found to a stored representation
         1. This is too black and white, and there is too much variation in speech production for this model to work.
         2. Dialect, production rate, loudness etc. – it’s not efficient.
      2. Analysis by synthesis – recognition speech by reference to the actions necessary to produce a sound.
         1. Basically that we internally produce speech sounds until it matches what we are hearing.
         2. This is biased by acoustic cues
         3. Uses our speech production skills to understand speech recognition
         4. Interspeaker differences are ok because we are generating our own candidates for the words
         5. Constraints based on plausible words
         6. Motor theory – the speech synthesizer models the articulation and motor movements of the speaker.
         7. Problems
            1. No way to translate the hypotheses into heard speech for a match (basically you don’t say it aloud)
            2. We are able to figure out improbable words that are spoken well, so it’s probably all data driven
      3. Cohort model – talks about bottom up nature (below)
      4. Connectionist TRACE model – interactive nature (below)
   2. Cohort model
      1. When we hear speech we set up a cohort of possible items that the word could be.
         1. Items are then eliminated from the set until one is left
         2. Originally this model was more interactive, but later versions were more autonomous.
      2. Stages of processing
         1. Access stage – perceptual representation is used to active possible cohorts
            1. Especially the beginnings of words, very important.
         2. Selection stage – when only one item is chosen.
         3. Integration stage – semantic and syntactic properties are activated (postlexical)
      3. The original model let all the other information (context) interact with eliminating the other members of the cohort and other members tried to block each other out (like the logogen model)
      4. Uniqueness point
         1. /t/ - loads of words
         2. /tr/ - still a lot
         3. /tre/ - trespass, trestle, trend, trench
         4. /tress/ - trespass, tress, trestle
         5. /tresp – uniqueness point
      5. Recognition point and uniqueness point do not have to match. In “the hunter ignored the sign not to tres-“ you might recognize before the uniqueness point
      6. The original version biased too much based on context and made errors. Also, if sensory information was degraded it was much slower than we normally are.
      7. Revised model
         1. Bottom – up priority – context cannot be used to restrict which items form in the initial cohort
         2. Context can’t be used to eliminate items until after the uniqueness point
         3. Also words are not completed eliminated from the cohort just moved down so that we can recognize mispronounced words or degraded (bleasant).
         4. Semantic information has a late effect to help with integration of the early winner
   3. Tests of cohort model
      1. Shadowing – listen to continuous speech and repeat it back as you go. The speech has mistakes in it.
         1. About half the time, participants report back the information as it should be, instead of how they heard it.
         2. Fluent restorations – (see above) such as saying tragedy when you heard travedy
         3. Most of the restorations were made when:
            1. The distortion was slight, distortion in the final syllable, word was highly predictable from context.
            2. Implies that semantics occurs very quickly, not after a clause.
            3. Also shows that bottom-up processing is more predominant.
      2. Listening for mispronunciations task
         1. Listen to distorted speech and look for the changes
         2. We are sensitive to problems at the beginnings of words
      3. Gating task – a task that involves revealing progressively more of the sound of the word
         1. Shows the effect of context – people need 333msec for words by themselves, but 200msec for words in context
         2. Visual gating task – shown pictures of different objects, then asked to click on the … at cl- you look at clown and cloud, but after the uniqueness point, you look at one over the other
         3. Cross Modal priming – listen to words and do a lexical decision task on the screen
            1. Priming for money and ship at “cap” (capitol, captain)
            2. After uniqueness point, only primes the related words (captain > ship). Supports the hypothesis that word selection is influenced by context late in the game.
   4. Lexical neighborhoods influence
      1. All possible words are analyzed in parallel, so large cohorts (competitors, neighborhoods) should not have an effect
      2. We are less able to identify words with high frequency neighbors – neighborhood density
      3. Words with many neighbors take longer to identify, especially when the neighbors are also high frequency
   5. The cohort model? In summary
      1. Early version – context didn’t affect access stage, but selection and integration
      2. Later version – context only affects integration
      3. Elements have activation levels based on input , candidates are analyzed in parallel, allows activations to decay
7. TRACE model
   1. Definition – highly interactive model of word recognition
   2. Characteristics
      1. Emphasizes the role of top down processing (context), so that lexical context can help with acoustic processing, and word level can influence word processing
      2. Connectionist model – many simple processing units strung together
      3. Three levels of processing
         1. Early acoustic processing – phonemic features
         2. Phonemes
         3. Output – word level
      4. Activation starts with features level and gets activated and ends with one output node winning as “recognized”
      5. Information flows up and down levels so that context can affect phonemic feature perception
      6. Inhibitory units eliminate competitors as they get activated
   3. Accounts for word initial effect – the beginnings of words are the most important to recognition because they start the activation in the right direction.
   4. Evaluation of trace
      1. Handles context well, can mimic the phonemic restoration effect, acoustic variability = ok
      2. Finds word boundaries and deals with noise well
      3. Problems
         1. Too many parameters, so it can model anything (does that make it realistic?)
         2. Time course slices are not realistic
         3. Some problems with probabilistic phoneme recognition
            1. L or R? s\_i, p\_i, t\_i
         4. Controversial that top down processes affect speech perception
         5. Trouble with mispronunciations – best word is activated then corrected on a phoneme level. But it slows down time (which we don’t do), because we have only one output in the model
         6. Trouble with nonword processing
      4. Outputs –
         1. Single output – only one answer is activated for recognition
         2. Multiple output – race model – lexical and prelexical, whichever one finishes processing first gets used, you can also bias processing to one side or other – causes problems for TRACE because single output.
8. Other connectionist models
   1. Recurrent connections – (figure 9.6 page 278) context information is stored in the hidden layer – or the last activations, helps them deal with time issues
   2. Updated update cohort model that is connectionist – mainly different from TRACE in using distributed units for all layers of the model (whereas TRACE maps phonemes to lexical units directly)
      1. Deals well with nonwords
      2. Allows for competition between words based on coherence (how alike sounds of the neighbors are in a cohort – captain > captive)
         1. However, the semantic properties will be unrelated
         2. More incoherent sets are, the more competition
   3. MERGE/SHORTLIST
      1. Both connectionist bottom up only models that are races between competing words
9. Comparison
   1. Hear speech > segment into words
      1. Models that use syllables will be slower because you need more of the word before the model starts (cohort)
      2. Other models that use stressed syllables have the disadvantage that they can’t recover from errors
         1. Revised cohort with variegated activation solves this problem.
      3. TRACE – doesn’t have these problems because sounds are mapped, but that can create large neighborhoods and be slow
         1. Top down activation is problematic, does it really inhibit neighbors?
      4. Problems for all models: we use other information to help us, such as lip reading.
10. Brains!
    1. Brain damage can causes lots of problems with prelexical code
    2. Vowels and consonants seemed to be processed by different systems
    3. Pure word deafness – can speak, read, write normally, but can’t understand speech (problem with prelexical processing)