Psych 215L: Language Acquisition

Lecture 3
Speech Perception

Learner’s job: Identify phonemes (contrastive sounds that signal a change in meaning)

Phonemes are language-specific - /r/ is a phonemic contrast in English but not in Japanese

Kids of the world require knowledge of phonemes before they can figure out what different words are - and when different meanings are signaled by different words

Important: Not all languages use the same contrastive sounds. Languages draw from a common set of sounds (which can be represented by the International Phonetic Alphabet (IPA)), but only use a subset of that common set.

Child’s task: Figure out what sounds their native language uses contrastively.

meaningful sounds in the language: “contrastive sounds” or phonemic contrasts

Acoustic
Innate
Phonemic
Constructed
Speech Perception: Computational Problem

Divide sounds into contrastive categories (phonemes)
Here, 23 acoustically-different sounds are clustered into 4 contrastive categories. Sounds within categories are perceived as being identical to each other.

---

Speech Perception: Computational Problem

Real world data are actually much harder than this…
(from Swingley 2009)

---

Categorical Perception

Categorical perception occurs when a range of stimuli that differ continuously are perceived as belonging to only a few categories with no degrees of difference within a given category.

Actual stimuli

Categorical Perception of stimuli

---

Acoustic-Level Information

Includes: timing and frequency
Tones: frequency (close-up)
Acoustic-Level Information

**Language sounds**

Vowels combine acoustic energy at a number of different frequencies.

Different vowels ([a] "ah", [i] "ee", [u] "oo" etc.) contain acoustic energy at different frequencies.

Listeners must perform a frequency analysis of vowels in order to identify them (Fourier Analysis).

---

Acoustic-Level Information

**Language sounds**

Male Vowels (close up)

---

Synthesized Speech

**Language sounds**

Female Vowels (close up)

---

Synthesized Speech

**Language sounds**

Valuable tool for investigating perception: [Praat](www.praat.org)
Acoustic-Level Information

Language sounds
Timing: Voicing

Acoustic-Level Information

Language sounds
Timing: Voice Onset Time (VOT)

English VOT production
Not uniform - there are 2 categories (distribution is bimodal)

Perception of stimuli: 2 categories
Perceiving VOT

*Categorical Perception*: 
\[
\begin{array}{c}
\text{dQ} \\
\text{tQ}
\end{array}
\]

Decision between d/t
Identification task: “Is this sound d or t?”

Discrimination Task
“Are these two sounds the same or different?”

Same/Different
0ms 60ms

Same/Different
0ms 10ms

Same/Different
40ms 40ms

Discrimination Task
“Are these two sounds the same or different?”

\[
\begin{array}{ccc}
D & 0\text{ms} & 20\text{ms} & D \\
D & 20\text{ms} & 40\text{ms} & T \\
T & 40\text{ms} & 60\text{ms} & T
\end{array}
\]

Why is this pair difficult?
(i) Acoustically similar?
(ii) Same Category?

Across-Category Discrimination is Easy
Within-Category Discrimination is Hard
Cross-language Differences

Identification task:
English speakers can discriminate r and l, and seem to show a similar pattern of categorical perception to what we saw for d vs. t

Cross-Language Differences

Discrimination task:
English speakers have higher performance at the r/l category boundary, where one sound is perceived as r and one sound is perceived as l. Japanese speakers generally perform poorly (at chance), no matter what sounds are compared because r and l are not contrastive for them.

Miyawaki et al. 1975

Hindi

dental [d]
(tip of tongue touches back of teeth)

retroflex [D]
(tongue curled so tip is behind alveolar ridge)

English [d] is usually somewhere between these
Cross-Language Differences

Salish (Native North American language):
- Glatalized voiceless stops
  - Uvular – tongue is raised against the velum
  - Velar – tongue is raised behind the velum

(they are actually ejectives - ejective is produced by obstructing the airflow by raising the back of the tongue against or behind the velum)

Perceiving sound contrasts

Key...
- This ability to distinguish sound contrasts extends to phonemic contrasts that are non-native. (Japanese infants can discriminate contrasts used in English but not in Japanese, like ri. This goes for both vowels and consonants.

...vs. adults
- Adults can’t, especially without training - even if the different is quite acoustically salient.

So when is this ability lost?
- And what changes from childhood to adulthood?

Speech Perception of Non-Native Sounds

Comparing perceptual ability

Werker et al. 1981: English-learning 6-8 month olds compared against English & Hindi adults on English & Hindi contrasts


But when after 6-8 months is the ability to lost?  Werker & Tees (1984)

Key into “critical period” hypothesis for language (Lenneberg 1967) - when language can be learned natively

“To test for this critical period, children of 12 and 8 years were tested, with the expectation that the 8-year-olds but not the 12-year-olds would be able to discriminate nonnative contrasts. English-speaking children of both ages, however, performed like English-speaking adults... study was extended to 4-year old children, who actually performed most poorly of all on nonnative contrasts... findings revealed that experience must begin to influence speech perception long before 4, certainly well before the critical period suggested by Lenneberg.”
Speech Perception of Non-Native Sounds

But when after 6-8 months is the ability to lost?           Werker & Tees (1984)

Salish & Hindi contrasts

Change happens somewhere around 8-10 months, depending on the sound contrast.
See Yoshida et al. (2010) for evidence that infants have some malleability still at 10 months, but it's much less than at 6 or 8 months.

More about contrastive sounds

There are a number of acoustically salient features for sounds. All it takes for sounds to be contrastive is for them to have "opposite" values for one feature.

Example:
English sounds "k" and "g" differ only with respect to voicing. They are pretty much identical on all other features. Many contrastive sounds in English use the voicing feature as the relevant feature of contrast (p/b, t/d, s/z, etc.). However, there are other features that are used as well (air flow, manner of articulation, etc.).

Task for the child: Figure out which features are used contrastively by the language. Contrastive sounds for the language will usually vary with respect to one of those features.

Discovering contrastive sounds: What's the point of it again?

The idea is that once children discover the meaningful sounds in their language, they can begin to figure out what the words are.

Ex: An English child will know that "cat" and "caat" are the same word (and should have the same meaning).

As adults, we can look at a language and figure out what the contrastive sounds are by looking at what changes a word's meaning. But children can't do this - they figure out the contrastive sounds before they figure out words and word meanings.

Experimental Study:
Dietrich, Swingley & Werker (2007)

Testing children’s perception of contrastive sounds

Dutch and English contrastive features differ.

In English, the length of the vowel is not contrastive

“cat” = “caat”

In Dutch, the length of the vowel is contrastive

“cat” ≠ “caat”

(Japanese also uses this feature)
Does the data distribution show this?
Dutch and English vowel sounds in the native language environment also seem to differ

“...studies suggest that differences between the long and short vowels of Dutch are larger than any analogous differences for English.”

Dutch = bimodal distribution?  
English = unimodal distribution?
Does the data distribution show this?

Dutch and English vowel sounds in the native language environment also seem to differ

“…studies suggest that differences between the long and short vowels of Dutch are larger than any analogous differences for English.”

![Graph showing frequency of sound in input vs. vowel duration for Dutch and English. Dutch is bimodal distribution, English is unimodal distribution.]

Learning from real data distributions

How do we know that children are sensitive to distributional information?

Maye, Werker, & Gerken (2002)

Created synthetic sounds ranging from [da] to [ta] that were non-native for the infants (because they were unaspirated).

Maye, Werker, & Gerken (2002)

- Familiarized 6 to 8-month-old infants to one of two sets
  - Bimodal Set: Sounds on the ends near [da] and [ta].
  - Unimodal Set: Sounds in the middle.
- Test preference for:
  - 3 6 3 6... (Alternating) vs. 3 3 3 3... (Non-alternating) stimuli
Maye, Werker, & Gerken (2002)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Condition</th>
<th>Median (IQR)</th>
<th>Alternating Trials (s)</th>
<th>Non-Alternating Trials (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 months</td>
<td>Unimodal</td>
<td>4.88 (0.47)</td>
<td>=</td>
<td>4.53 (0.51)</td>
</tr>
<tr>
<td></td>
<td>Bimodal</td>
<td>5.66 (0.44)</td>
<td>=</td>
<td>6.41 (0.32)</td>
</tr>
<tr>
<td>8 months</td>
<td>Unimodal</td>
<td>4.98 (0.63)</td>
<td>=</td>
<td>5.20 (0.56)</td>
</tr>
<tr>
<td></td>
<td>Bimodal</td>
<td>5.45 (0.52)</td>
<td>&lt;</td>
<td>6.15 (0.56)</td>
</tr>
</tbody>
</table>

Infants trained on the Bimodal data had a novelty preference for non-alternating trials. They learned to expect alteration, and were surprised by non-alteration.

Maye, Werker, & Gerken (2002)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Condition</th>
<th>Median (IQR)</th>
<th>Alternating Trials (s)</th>
<th>Non-Alternating Trials (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 months</td>
<td>Unimodal</td>
<td>4.88 (0.47)</td>
<td>=</td>
<td>4.53 (0.51)</td>
</tr>
<tr>
<td></td>
<td>Bimodal</td>
<td>5.66 (0.44)</td>
<td>=</td>
<td>6.41 (0.32)</td>
</tr>
<tr>
<td>8 months</td>
<td>Unimodal</td>
<td>4.98 (0.63)</td>
<td>=</td>
<td>5.20 (0.56)</td>
</tr>
<tr>
<td></td>
<td>Bimodal</td>
<td>5.45 (0.52)</td>
<td>&lt;</td>
<td>6.15 (0.56)</td>
</tr>
</tbody>
</table>

Infants trained on the Unimodal data did not prefer/disprefer one over the other. They did not seem to learn any expectation.

Maye, Werker, & Gerken (2002)

Back to Dietrich, Swingley, & Werker (2007)

Dutch and English vowel sounds in the native language environment also seem to differ

"...studies suggest that differences between the long and short vowels of Dutch are larger than any analogous differences for English."

<table>
<thead>
<tr>
<th>Frequency of sound in input</th>
<th>Dutch = bimodal distribution?</th>
<th>English = unimodal distribution?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dutch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>English</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Back to Dietrich, Swingley, & Werker (2007)

Prediction if children are sensitive to this distribution

Dutch children interpret vowel duration as a meaningful contrast because the distribution is more bimodal

Implication: Change to vowel duration = new word

English children should not interpret vowel duration as a meaningful contrast because the distribution is more unimodal

Implication: Change to vowel duration = same word as before

Dietrich, Swingley, & Werker (2007)

Tests with 18-month-old children who know some words (and so have figured out the meaningful sounds in their language)

"Switch" Procedure: measures looking time

Habituation

Test

Dietrich, Swingley, & Werker (2007)

Experiment 1: Testing English and Dutch kids on Dutch vowel durations

Test

Dutch kids

5.04 sec

9.23 sec

difference

English kids

6.66 sec

7.15 sec

no difference

Frequency of sound in input

Vowel duration

Dietrich, Swingley, & Werker (2007)

Experiment 2: Testing English and Dutch kids on English vowel durations

Test

Dutch kids

5.92 sec

8.16 sec

difference

English kids

7.34 sec

8.04 sec

no difference

Frequency of sound in input

Vowel duration
**Dietrich, Swingley, & Werker (2007)**

**Experiment 3:** Testing English and Dutch kids on vowel quality contrast (a/e)

<table>
<thead>
<tr>
<th>Frequency of sound in input</th>
<th>Vowel duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test</td>
<td>0</td>
</tr>
<tr>
<td>Dutch kids</td>
<td>4.08 sec</td>
</tr>
<tr>
<td>English kids</td>
<td>6.31 sec</td>
</tr>
<tr>
<td>Difference</td>
<td>2.23 sec</td>
</tr>
</tbody>
</table>

(This is a control condition to make sure English kids can do the task when the sound is contrastive for them)

**What drives children to learn the distinction?**

“One frequently raised hypothesis...is that it is driven by contrast in the vocabulary. Dutch children might learn that [a] and [e:] are different because the words [stat]...and [sta:t]...mean different things...however, children that young do not seem to know many word pairs that could clearly indicate a distinction between [a] and [e:].”

**Implications of experiments 1, 2, and 3:** Dutch children recognize vowel duration as contrastive for their language while English children do not. This can only be due to the data encountered by each set of children in their language.

**Just a note that experimental data with infants is messier than it sounds.**
Dietrich, Swingley, & Werker (2007)

“The other current hypothesis is that children begin to induce phonological categories “bottom-up”, based on their discovery of clusters of speech sounds in phonetic space…undoubtedly implicated in infants’ early phonetic category learning, which begins before infants know enough words for vocabulary-based hypotheses to be feasible…”

“A necessary condition for such learning to be the driving force behind Dutch children’s phonological interpretation in the present studies is that long and short vowels be more clearly separable in Dutch than in English… preliminary examination of this problem using corpora of Dutch child-directed speech indicated that the set of long and short instances formed largely overlapping distributions.”

Implication: Dutch children need other cues to help them out.

Swingley (2009)

One potential source of information: keep some contextual information for each vowel sound (what word it came from, if it comes from a frequent word).

Vallabha et al. (2007)

Also, not all distributions (and categorical features) may be so difficult to extract from acoustic information alone (like F1, F2, and duration).

English vs. Japanese on 3 acoustic dimensions, from child-directed speech:
- F1, F2, and Duration
- English: F1 vs F2 creates 4 vowels
- Japanese: F1 vs duration creates 4 vowels
**Formants**

F1: depends on whether the sound is more open or closed. (Varies along y axis.) F1 increases as the vowel becomes more open and decreases as vowel closes.

F2: depends on whether the sound is made in the front or the back of the vocal cavity. (Varies along x axis). F2 increases the more forward the sound is.

Idea: As long as speakers use the same values for these formants, they will produce the same vowel.

---

**Vallabha et al. (2007)**

A model trained on child-directed speech data can (mostly) find the four vowels appropriate for each language.

---

**An issue**

An issue: There is considerable variation in formants (like F1 and F2) between speakers. How can they get these speaker-neutral values for these features?

---

**Monahan & Idsardi (2010)**

Human brains may be biased to extract this information by using certain normalization procedures.

“...We propose a novel formant ratio algorithm in which the first (F1) and second (F2) formants are compared against the third formant (F3). Results from two magnetoencephalographic experiments are presented that suggest auditory cortex is sensitive to formant ratios...we present statistical evidence that this algorithm eliminates speaker-dependent variation based on age and gender from vowel productions...”