Distributions of individual segments and of phonological features

Emily Moeng | moeng@email.unc.edu | www.unc.edu/~moeng

The Distributional Hypothesis

- How do infants acquire early phoneme categories at such an early age?
- **The Distributional Hypothesis**: Claims that infants acquire early phoneme categories by attempting to frequency-based distributions in the ambient language

![Diagram of token frequency and number of frequency peaks](image)

**Problem**: Actual distributions in natural language are problematic for the Distributional Hypothesis

- Although influential, natural language exhibits large category overlap, which is problematic for the Distributional Hypothesis
- **Example**: English vowel qualities (Swingley, 2009)

**Research Questions**

1. How strong is the “category overlap problem” for different types of phonemes? Is category overlap only seen in phonemes which do not have “anatomical landmarks” (i.e., vowels/glides)?
2. Does collapsing segments by feature aid machine learners in learning categories?

**Methodology**

- Data collected from the infant-directed speech of a single male speaker in French using the CHILDES corpus (MacWhinney, 2000; Yamauchi, 2007)
- Distributional information was measured by testing how well Gaussian learning algorithms correctly classified phonemes (following the methodology of Adriaans and Swingley, 2013)
- Used the Expectation-Maximization algorithm (MCLUST package in R, Fraley and Raftery, 2006) to estimate a purely Distributional learner
- Given no information as to the number of desired clusters
- Compared the distributions of the following (all are separate phonemes in French):
  - Stops /k/ /g/ (Voice Onset Time)
  - Fricatives /s/ /ʃ/ (Center of Gravity)
  - Glides /j/ /w/ /ɥ/ (F2 and F3)
  - Vowels /i/ /y/ /u/ (F2 and F3)

**Results: Actual vs. Learned Categories**

![Graphs showing actual vs. learned categories](image)

**2 Metrics to Evaluate Results**

- **Metric 1: Classification Accuracy**
  - What percent of each phoneme was placed into the correct “learned phoneme” category
  - [Figure showing classification accuracy](image)

- **Metric 2: Psychophysical Distance Accuracy**
  - How far from the actual mean is the learned mean in perceptual space?
  - [Figure showing psychophysical distance accuracy](image)

**Conclusion and Discussion**

**Table 1**: Average and standard deviation of learned means for formants (F1, F2, and F3)

- Pairs of formants were found that contained phonemes and formants
  - Frequency overlap was found to occur in all phoneme types tested
  - Formants were used to measure the perceptual distance, rather than the acoustic distance

**Selected References**