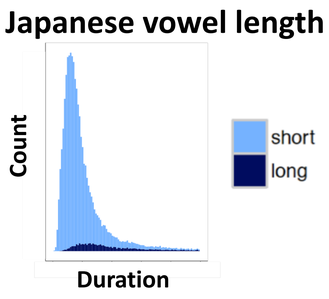
Naturalistic data support distributional learning across contexts

Kasia Hitczenko[1] and Naomi H. Feldman[2,3]

1. Northwestern Linguistics, 2. University of Maryland Linguistics, 3. University of Maryland Institute for Advanced Computer Studies

Background

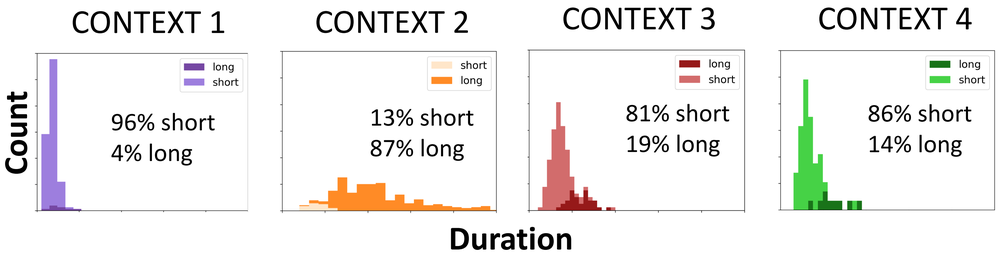
- **Distributional Learning:** infants learn that sounds are contrastive from bimodal distributions [1]
- Not supported by naturalistic input [2]



Distributional Learning Across Contexts

Proposal: Infants learn a dimension is contrastive, by tracking its distribution shape across contexts, and noticing that the shape varies across contexts

Why do we expect the distribution shape along a dimension to vary across contexts if it's contrastive?



Is the necessary signal present in the data?

Data & Methods

Data = naturalistic speech corpora

- French vs. Japanese (no length contrast vs. length contrast) [NCCFr [3] vs CSJ [4]]
- Dutch (vowels w/ length contrast vs. w/o) [ECSD [5]; Fikkert/Levelt/Swingley IDS [6,7,8]]

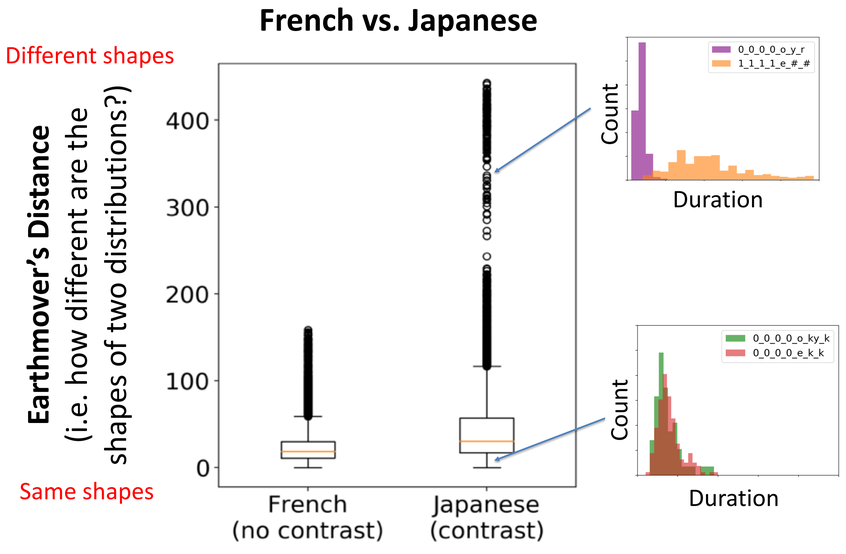
Two ways of defining contexts:

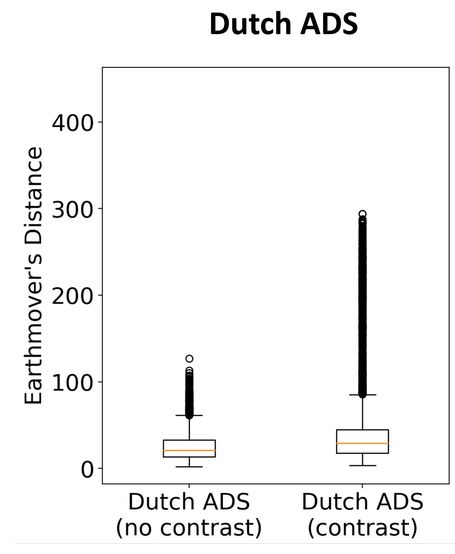
1. Prosodic position + vowel quality + neighboring sounds
2. Word frames

Methods

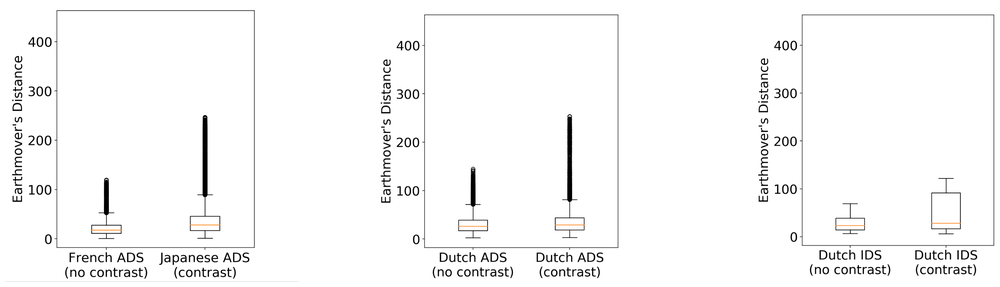
1. Get distribution along duration for each context
2. Compare them pairwise using Earthmover's distance [9]

Results





Also holds when word frames are the contexts:



1. The distribution shape along a dimension varies more across contexts when the dimension is contrastive than when it's not
2. This is signal that infants could use to learn which acoustic dimensions are contrastive in their language
3. One of the first distributional learning accounts supported by naturalistic data

Citations: [1] Maye et al. (2002); [2] Bion et al. (2013); [3] Torreira et al. (2010); [4] Maekawa (2003); [5] Ernestus (2000); [6] Swingley (2019); [7] Levelt (1994); [8] Fikkert (1994); [9] Rubner et al. (1998)

Funding: NSF IIS-1421695, NSF BCS-1734245, NSF NRT #1449815

Acknowledgements: Thank you to T. Schatz, M. Ernestus, D. Swingley, & F. Torreira for help acquiring corpora.