A Simple Model of Infant Word Finding Using Computational Phonetics Embeddings

Taiwo Adeaga (COL 2026), Ziana Sundrani (COL 2027)
Mentor: Dr. Daniel Swingley, School of Arts and Sciences, Department of Psychology
Funded by the Penn Undergraduate Research Mentoring Program (PURM)

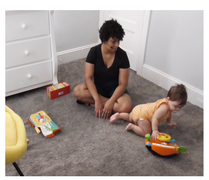
Introduction

- Babies use word and pattern recognition to build their vocabulary and hone their sound perception
- However, when children learning language hear sentences, word boundaries are often unclear
    - Words sound somewhat different every time they're pronounced, which makes identifying them before knowing the language harder
- So, how do babies recognize words?
- Can infants learn from portions of sentences that repeat? Do repeating segments correlate with words or common phrases?
- Our goal was to learn how babies use infant-directed speech to build their lexicon. We used speech technology to make a model of a baby using this strategy and evaluate whether the repeated sequences resembled words

Methods

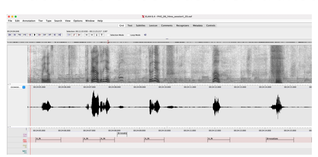
Step 1: Data Collection

- A mother was set up with a mic in her home and left to naturally to her 7-month-old
    - This recording comes from a larger data set of maternal infant-directed speech




Step 2: Transcription

- Maternal infant-directed utterances were isolated from the original recording
- Each utterance was then transcribed word for word

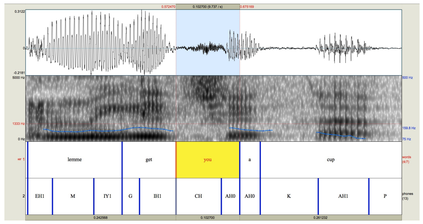


```
OK
Ooh look
You made a smelly smelly ?
What ?
Uh we put them on backwards uhoh
Smelly smelly Kayla bear
You got a white thing and you got a pink pig
```

Methods (continued)

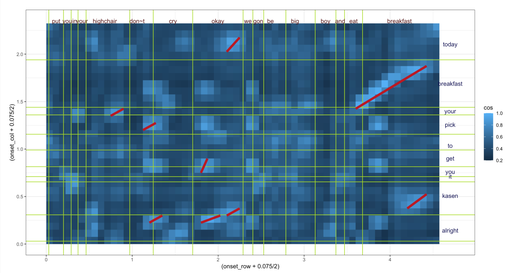
Step 3: Word and Phone Alignment

- Transcriptions and their corresponding audios were processed through the Montreal Forced Aligner (MFA) software to create files of aligned phones and words to the audio file (McAuliffe et al., 2017)
- Native English speakers corrected MFA outputs to more accurately reflect the sounds spoken and timestamps of the words and phones



Step 4: Embeddings

- The utterances were characterized into a mathematical representation of speech, in the form of a 512-dimensional vector, called an embedding (Swingley & Algayres, 2024)
- The sound representations of nearby utterances were compared to each other and their similarity was computed using cosines (Park & Glass, 2008)



One utterance is on the x-axis and the other is on the y-axis. The first sound of the first utterance is compared to every sound of the second sentence. This is done for all sounds of both sentences. The stretches of light blue boxes mean that there was a stretch of time in which the sounds of both sentences were very similar.

- These segments of similar sounds were the hypotheses of a word a baby would pick out

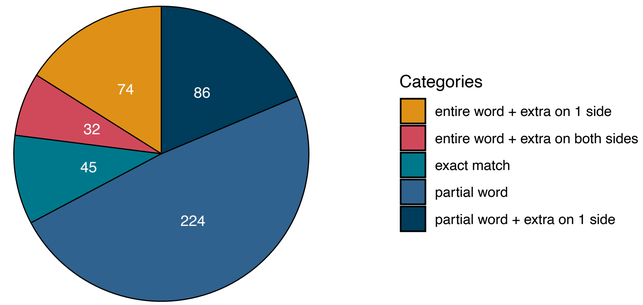
Step 5: Evaluation

- Do the segments found in step 4 resemble real English words?
- To answer these questions, each segment identified in Step 4 was isolated and compared to a "gold standard" word determined by the transcriptions from Step 2

Results

To analyze the accuracy of this approach to building a vocabulary, the following categories were created:

- Exact match: Y EY = Y EY ("yay")
- Partial word: AH M is only part of S AH M ("some")
- Partial word + extra on 1/both sides: T IH S W AO has part of the target word P R AE K T IH S ("practice") but also has sounds from W AO K IH NG ("walking")
- Entire word + extra on 1/both sides: AH T Y UW G AA contains the full target word Y UW ("you") but also has sounds from AH T ("what") and G AA ("got")



Discussion

- Some words were able to be identified through this word finding strategy, however, a majority of segments were not exact words
- About half the segments were only part of a word and were missing phones at the beginning and/or end of the segment
- This may mean our mathematical speech representations did not do well with managing coarticulation, which occurs when the last sound of the previous word bleeds into the first sound of the following word. Thus, the embeddings may not have been as good as humans at recognizing word boundaries because similar patterns of sounds were surrounded by different phonetic environments.
- To further understand this data set, we want to compare our results to a chance model that randomly creates segments to determine if our model performs better
- We also plan to do further analysis on the non-exact matches to analyze how different they were from their target words in the hopes of understanding other strategies babies may use to learn language

References

- McAuliffe, Michael & Socolof, Michaela & Mihuc, Sarah & Wagner, Michael & Sonderegger, Morgan. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. 498-502.
- Park, A. S., & Glass, J. R. (2008). Unsupervised pattern discovery in speech. *IEEE Transactions on Audio Speech and Language Processing*, *16*, 186–197.
- Swingley, D., & Algayres, R. (2024). Computational modeling of the segmentation of sentence stimuli from an infant Word-Finding study. *Cognitive Science*, *48* , e13427.