

Emergence of phonemic categories through the acquisition of a lexicon

Annika Heuser¹

¹*University of Pennsylvania*

Introduction: Infants must learn to map continuous speech signals to discrete phonemic categories to learn the patterns of their language. By 6 months, infants are already perceptually tuning to their language’s sound categories [1] and losing the ability to differentiate non-native contrasts by 12 months [2]. However, 9-month-olds can regain the ability to differentiate a non-native contrast if it is consistently lexically contrastive [3], highlighting the role of word learning in phonemic category acquisition. We posit that rather than grouping together segments extracted from different words, the infant’s goal is to group together acoustic realizations of the same word [4-5]. The infant need only know that different words have different meanings, rather than the precise semantics. Because words are made up of different phonemes, divisions between phonemes may arise from generalizable divisions between words.

Model: We built a computational model representative of this theory, inspired by the model in [5]. We sampled monosyllabic words from TIMIT and the Philadelphia Home Corpus (PHC, 17 transcribed sessions of 11 mothers talking to their 6-14-month-old infants), respectively. The model represents words from TIMIT as 20-dimensional exemplars, where each dimension corresponds to acoustic differences that infants can perceive (e.g. [1], [6]). Words from PHC are represented as 16-dimensional exemplars, excluding the closure-related features (Table 1). If the model learns a new word that it places in the same region of its 16- or 20-dimensional perceptual space as another acquired word, then it determines the relative cost of this homophone representation vs. a contrastive representation. If the contrastive representation is less costly, then it places a max margin classifier boundary between the two words (Fig. 1).

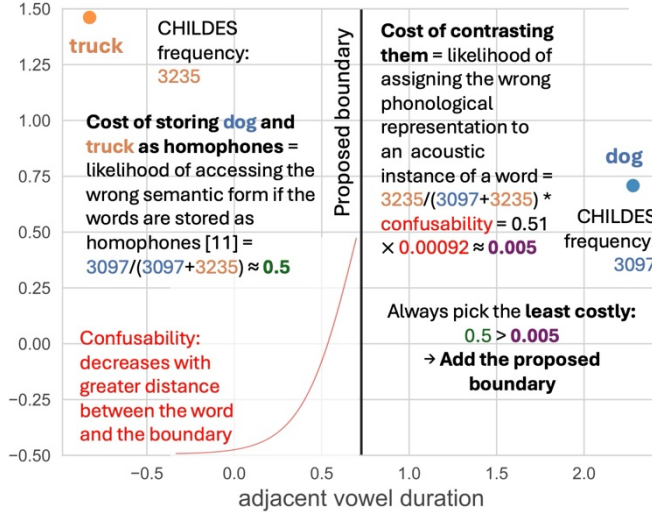
Method: We evaluated the model via the ABX task, where A and X are different acoustic instances of the same word (e.g. “dog”) while B is an acoustic instance of a minimal pair word (e.g. “dig”). The model is successful when there is a smaller edit distance between the classifications (i.e. the 0s and 1s encoding on which side of each of the model’s boundaries the word instance is situated) of A and X than between B and X. We compared the TIMIT and PHC models’ ABX performance to that of three baselines, and to that of unsupervised models trained on 50-400 hours of speech, reported in [7]. The three baselines consist of 1) the same number of decision boundaries as the model but randomly placed in perceptual space, and 2) k-means (k=50) and 3) gaussian mixture (GMM; components=50) models trained on individual segments. The purpose of the first baseline was to verify that our model does not only succeed because of the number of boundaries, nor the features defining the perceptual space itself. The other baselines and the models in [7] represent the strategy of distributionally learning across units smaller than words. We ran 10 training/testing iterations for our model and each baseline.

Results: Phonemic categories appear to solidify when the model has acquired ~50 words (the expected size of a 12-month-old’s lexicon [4]). The TIMIT and PHC models outperform the baselines and the models trained on ≤ 100 hours of speech in [7] (Fig. 2). The number of boundaries our model adds to keep acquired words apart tapers off, indicating that existing boundaries generalize well (Fig. 3). Evidence that 6–12-month-old infants can segment words that they know from continuous speech (e.g. [8]) helps justify the model assumption of relatively consistent segmentation of acquired words. However, infants may also learn from imperfectly segmented words, so we tested the model on TIMIT/PHC data where the word and phone segmentation boundaries were stochastically perturbed up to 50% of the time. We found the model’s performance to be robust, indicating that neither perfect word segmentation nor perfect feature extraction is crucial to the model’s success (Fig. 4). Its success not only supports our theory that phonemic category acquisition is driven by lexical development, the model itself constitutes a formal mechanism for this theory. We thereby contribute a better understanding of how infants might learn to map speech to discrete phonemic categories.

Table 1: Example TIMIT word exemplars. The features in pink rows are not included in PHC word exemplars. The model would use the purple feature to separate these words.

	dog	truck
onset duration	-0.727	1.408
onset closure duration	0.191	1.055
onset closure partial voice	0.070	-0.189
onset other partial voice	-0.570	-0.684
onset transition f0	0.830	-0.341
onset transition f1	0.064	0.913
onset silence	-0.366	-0.430
adjacent vowel duration	2.280	-0.827
mean f0	0.301	-0.128
mean f1	0.708	1.461
mean f2	-0.823	-1.461
coda duration	-0.573	1.310
coda closure duration	1.505	-0.275
coda closure partial voice	1.498	2.041
coda other partial voice	-0.222	0.991
coda transition f0	0.115	-0.136
coda transition f1	-0.142	0.534
coda transition f2	-0.811	-1.424
coda silence	0.118	-0.198

Fig. 1: Schematic illustration of the model decision of how to represent a new word in perceptual space.



References: [1] Kuhl, et al. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*.

[2] Werker & Tees (1984). Cross-language speech perception: Evidence....

[3] Yeung & Werker (2009). Learning words' sounds before learning how words sound: 9-month-olds use distinct objects as cues to categorize speech information. *Cognition*. [4] Swingley (2009). Contributions of infant word learning to language development. *Philosophical Transactions of the Royal Society B: Biological Sciences*. [5] Cui (2020). The emergence of phonological categories. University of Pennsylvania. [6] Aslin, et al. (1981). Discrimination of voice onset time by human infants: New findings.... *Child development*. [7] Lavechin, et al. (2025). Simulating early phonetic and word learning without linguistic categories. *Developmental Science*. [8] Bortfeld, et al. (2005). Mommy and me: Familiar names.... *Psychological science*. [9] Jusczyk & Aslin (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology*. [10] Van Orden (1987). A ROWS is a ROSE: Spelling, sound, and reading. *Memory & cognition*.

Fig 2: ABX performance of our models (trained on TIMIT and PHC) as they acquire more words. Our models and baselines were tested on unseen word types from TIMIT.

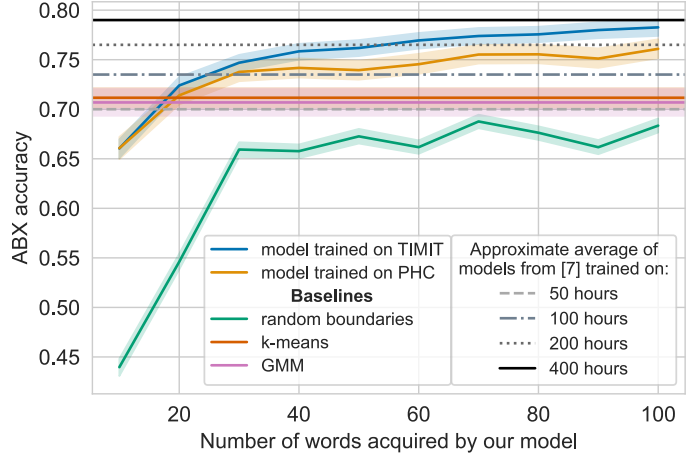


Fig. 3: The model adds fewer new boundaries as it acquires more words, and places a greater proportion of test words in regions unoccupied by training word exemplars.

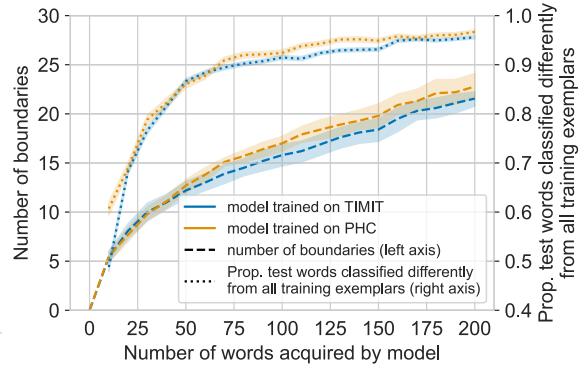


Fig. 4: When trained on perturbed data (acquired words = 100), our models continued to perform well on unperturbed test data of unseen word types from TIMIT. The baselines here for comparison were all trained on unperturbed data.

