Computer modeling suggests patterns of perceptual availability of phonological structure during infant language acquisition

Distinctive features like [±voice] and [±sonorant] have been a core construct of phonological theory for many decades [42, 24, 7, 8], and psycholinguistic evidence suggests that they are cognitively available to both adults [6] and infants [26, 21, 45]. Nonetheless, distinctive features are not directly observed by humans: they are abstractions that must be inferred from dense perceptual information (sound waves) during language acquisition and comprehension, which raises questions about how they are learned and recognized. Recent work on child language acquisition has stressed the importance of top-down (e.g. lexical and phonotactic) information for acquiring phonemic categories [33, 40, 16, 27, 31, 17, 11]. But to prevent the acquisition process from being circular, the acoustic signal must also provide evidence for phonemic categories. Furthermore, top-down guidance is likely less reliable to young infants, who must therefore rely more heavily on bottom-up perceptual information. To a learner faced with the immense challenge of discovering structure in dense perceptual input, do theory-driven phonological features “stand out” or are they swamped by noise?

We address this question through computational acquisition modeling, which permits fine-grained analysis of the learned representations that is not possible to obtain from human infants. Our acquisition model takes as a starting point cognitive evidence that brains actively model their perceptual world [15, 39, 47], that autoassociation characterizes the behavior of many brain regions [41, 37], that language comprehension and production might be linked through a sensorimotor loop [23, 14, 44, 46, 36, 25, 4], that limited auditory memory requires austere compression of dense acoustic percepts during real-time language comprehension [3, 2, 13], that featural decomposition of phone segments occurs during the acquisition process [26, 21, 45], and that there are broad tendencies toward categorical perception in human cognition [19], including that of infants [12]. The computational learners used in this study have all of these characteristics: they are deep neural autoencoders (percept modeling, autoassociation, sensorimotor loop) that force acoustic information from phone segments through a tight 8-dimensional representational bottleneck (compression) consisting of discrete binary stochastic neurons or BSNs (feature decomposition, categorical perception). Our learners thus have a representational capacity of 256 discrete phone categories, decomposable along 8 binary feature dimensions, with which to describe their variegated perceptual world. This setup allows us to evaluate degrees of correspondence between these perceptually-driven unsupervised representations and theory-driven phonological representations.

We deploy these models to answer two questions about the data available to young learners whose training signal must primarily be extracted from bottom-up perceptual information: (1) to what extent can phoneme categories emerge from a drive to model auditory percepts, and (2) how perceptually available are theory-driven phonological features? We apply our models to naturally-occurring acoustic phone segments from two typologically unrelated languages: Xitsonga and English [43]. Unsupervised phone classification metrics homogeneity (H), completeness (C), and V-measure (V) [38] are given in Tables 1a & 1b. As shown, much phonemic structure is perceptually available from acoustics alone (20-40x clustering improvement over a random baseline). We further analyze the recoverability of theory-driven phonological features from the learners’ latent bit patterns, using random forest classifiers [32] to fit propositional logical statements that map from latent bits to binary featurizations of the true segment labels [20, 18]. Precision (P), recall (R), and F-scores (F) are given in Tables 1c & 1d. Patterns of feature availability are remarkably consistent across languages, suggesting that the models are capturing generalized perceptual patterns. Furthermore, there are strong asymmetries in perceptual availability, with good recovery of voicing features and features that distinguish prototypical consonants from prototypical vowels, along with comparatively poorer recovery of e.g. certain place and manner distinctions. These findings align with attested patterns of infant phone discrimination [1, 10, 30, 34, 29, 5, 28, 35, 9, 22].

Our results show (1) that phonemic structures emerge naturally but imperfectly from perceptual reconstruction and (2) that theory-driven features differ in degree of perceptual availability. Together, these findings suggest that reliable cues to phonemic structure are immediately available to infants from bottom-up perceptual characteristics alone, but that these cues must eventually be supplemented by top-down lexical and phonotactic information to achieve adult-like phone discrimination. Our results also suggest fine-grained differences in degree of perceptual availability between features, yielding testable predictions as to which features might depend more or less heavily on top-down cues during child language acquisition.