Bottom-up learning of a phonetic system using an autoencoder

Frank Lihui Tan & Youngah Do
The University of Hong Kong
tt115889@connect.hku.hk, youngah@hku.hk

Human learners tend to perceive and acquire sound input categorically, even if the input is of a gradient nature [1, 2, 3]. This tendency results in the acquisition of a "prototypical phonetic system" [4], in which sounds near phonetic category centers are perceived as closer to their centers than they actually are. It is unclear how much of this prototypical system is learned through language experience and how much can be attributed to humans' innate cognitive properties, as infants as young as one month old already show categorical perception [1].

The current study aims to investigate whether 'naïve' learners can still acquire a prototypical phonetic system at the very initial stage of phonetic learning, before they have acquired any knowledge about their language's lexicon or structure. To answer this question, we test an autoencoder model that is trained purely in a bottom-up fashion, without assuming any abstract featural system. Unlike previous models, we evaluate the model's learning outcome directly on the hidden representations of phones, instead of on the basis of distinctive features.

Our results show that the autoencoder model is able to capture the phonetic system in the form of phone distributions in the hidden space, which is a close analogy to the prototypical category learning observed in human learners [4]. We selected two typologically unrelated languages, English and Mandarin, and used 38 hours of recordings from 40 English speakers (Buckeye Speech Corpus [5]) and a comparable size of recordings from Mandarin speakers (AISHELL-3 [6]) as training data. The recordings were in wave format, and we extracted mel-frequency cepstral coefficients before training. We then randomly segmented the data to ensure that no or minimal phonological cues and segmental boundary information was provided. No ground truth labels were incorporated. For each language, we built an autoencoder to encode the input information into a hidden space and reconstruct the hidden representation back to the input with least distortion [7]. The encoder simulated the complex, layered neural transformations underlying speech perception, which ultimately converted the sensory receptor signal to the underlying neural code of segments [8]. The decoder simulated the reverse process of generating sounds from internal representations, but excluded articulation since there is no simulation of articulators. The training was unsupervised, without external feedback, and no segment boundary was provided, which is analogous to the early stage of infants' phonetic acquisition [9].

The preliminary clustering task revealed that phonetic knowledge successfully emerged in the hidden space for both English and Mandarin languages. The models' hidden representations yielded significantly higher homogeneity, completeness, and V-measure scores than random clustering (e.g., $V_{\_Random\_English}$=0.006 vs. $V_{\_English}$=0.289), indicating that the autoencoder was able to reproduce the input sounds and identify phonetic category centers, even without phonological context or segmental boundary information. Further evaluations of the hidden representations showed that the model was able to project tokens of the same phone to similar areas and tokens of different phones to different areas in the hidden space, while successfully learning feature-based contrasts such as [±back], [±high], [±strident], and [±voice]. The current model trained solely on phonetic cues was able to construct a phonetic system that distinguishes sounds, implying that the model was able to project an acoustic token to its correct absolute position, rather than merely achieving paired phone contrasts.

Although the model achieved significant success, it did not capture the acquisition of allophones. Unlike top-down models [10, 11], which take into account human's different perceptual sensitivity and learnability of phonemes and allophones [12, 13], the current bottom-up model did not show a significant difference between phoneme and allophone projections. For example, the distribution of allophones of Mandarin /i/ (/i, ɪ˧, ɻ̩/) was similar to that of phonemes (μ (i,ɪ,ɻ dists)=2.978; μ (y, ɚ, a, ɤ, u dists)=3.194; p=0.817).

In summary, this study demonstrates that an autoencoder model can learn phonetic knowledge from contextless acoustic input without supervision or explicit segment boundary. The model was able to project different phones to different areas in the hidden space, similar to the way human infants acquire phonetic knowledge. This suggests that infants' phonetic knowledge may not be innate but can be acquired based purely on acoustic information, without relying on language-specific learning facilities. However, the model could not reach phonological knowledge without training on phonological cues, which is acquired by infants at around 8-10 months [14]. This implies that phonological information plays an indispensable role in the language-specific refinement of learners' knowledge on phonetics and phonology.

References

[1] Eimas, P. D. et al. (1971). Speech Perception in Infants. *Science*, *171*(3968), 303–306.

[2] Fernald, A., & Kuhl, P. (1987). Acoustic determinants of infant preference for motherese speech. *Infant behavior and development*, *10*(3), 279-293.

[3] Werker, J. F., & Lalonde, C. E. (1988). Cross-language speech perception: Initial capabilities and developmental change. *Developmental Psychology*, *24*(5), 672–683.

[4] Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics*, *50*(2), 93–107.

[5] Pitt, M. A. et al. (2007). *Buckeye Corpus of Conversational Speech (2nd release)*.

[6] Shi, Y. et al. (2021). *AISHELL-3: A Multi-speaker Mandarin TTS Corpus and the Baselines* (arXiv:2010.11567).

[7] Bank, D. et al. (2021). *Autoencoders*.

[8] Eggermont, J. J. (2001). Between sound and perception: reviewing the search for a neural code. *Hearing Research*, *157*(1–2), 1–42.

[9] Räsänen, O. (2014). Basic cuts revisited: Temporal segmentation of speech into phone-like units with statistical learning at a pre-linguistic level. *Annu. COGSCI*.

[10] Kolachina, S., & Magyar, L. (2019). What do phone embeddings learn about Phonology? *Proceedings of SSIGMORPHON*, 160–169.

[11] Silfverberg, M., Mao, L. J., & Hulden, M. (2018). Sound analogies with phoneme embeddings. *Proceedings of SCiL*, 136–144.

[12] Martin, A. et al. (2013). Learning Phonemes With a Proto-Lexicon. *Cognitive Science*, *37*(1), 103–124.

[13] Peperkamp et al. (2006). The acquisition of allophonic rules: Statistical learning with linguistic constraints. *Cognition*, *101*(3), B31–B41.

[14] Hayes, B. (2004). Phonological acquisition in Optimality Theory: The early stages. In R. Kager et al. (Eds.), *Constraints in Phonological Acquisition* (pp. 158–203). CUP.