Using Machine Learning to Build an Inventory of Sign Components

Lee Kezar[1], Jesse Thomason[1], Naomi Caselli[2], Zed Sehyr[3]
[1] *University of Southern California,* [2] *Boston University,* [3] *Chapman University*

Introduction. Like speech, signing contains more articulatory signal than is necessary to recognize the signs (e.g. size of the signing space, speed, repetition). But exactly how articulatory variation maps to the hypothesized abstract representational units of signs (phonology) remains an open theoretical question (Cann et al, 2012). This study introduces a computational method which identifies *spatiotemporal patterns* in the American Sign Language (ASL) lexicon, which we hypothesize to correlate with phonological features in ASL-LEX 2.0 (Sehyr et al, 2021). The byproduct of this method is *an inventory* of *sign components* which, in the context of the model, can be automatically and consistently spotted in video data. The inventory and model provide a data-driven platform for testing a number of hypotheses related to articulatory variation (e.g. accent, dialect) and the phonetics-phonology interface.

Method. Utilizing a machine learning model based on neural discrete representation learning[1] (van den Oord, 2017), we analyzed 91,000 videos of isolated signs from deaf early-exposed ASL signers gathered as part of another experiment (Kezar et al., 2023). The model was trained to generate eight "codes" for each sign, where each code represents a learned spatiotemporal pattern in the lexicon. The model's performance was evaluated by (a) its consistency in identifying signs and (b) the alignment of these codes with phonological features in ASL-LEX (Sehyr et al., 2021).

Results. The analysis showed the model could consistently identify signs with minimal confusion, averaging 1.3 unique code combinations among the videos of one sign (best case is 1.0, worst is 10.5). A significant correlation (Pearson's $r = 0.23$, $p < 0.0001$) between the codes and phonological features suggests a good alignment with ASL's phonological structure. Follow-up probing experiments revealed that some codes captured interpretable spatiotemporal patterns such as repeated movement and finger spread.

Conclusion. Phonological theorizing in sign languages has been hindered by the lack of conventional writing systems and the reliance on manual annotation of video data. This study demonstrates a promising data-driven approach for automatically building an inventory of sign components despite articulatory variability, highlighting the potential for these methods to inform sign language phonetics–phonology interface and further refine phonological theory in general (Brentari, 1999; Sandler & Lillo-Martin, 2006). For example, future extensions can study variation across and within sign languages by adjusting the language training data and granularity of the codes. Such an extension could lead to a more universal phonological inventory for sign languages and systematically accounting for variation from accents or dialects. Ultimately, tools for automated phonological transcription have played a pivotal role in advancing speech recognition technologies, and this approach could be used for similar purposes for SLs.

---

[1] *Representation learning* refers to methods where an artificial neural network's *hidden representation* (a vector of floating-point numbers generated for each input) is manipulated to exhibit certain properties. In van den Oord (2017), the network (an *encoder*) is constrained to represent any input as a sequence of *integers* [1, *N*] called "codes" such that a separate network (a *decoder*) can accurately reconstruct the input from the codes (input = output).

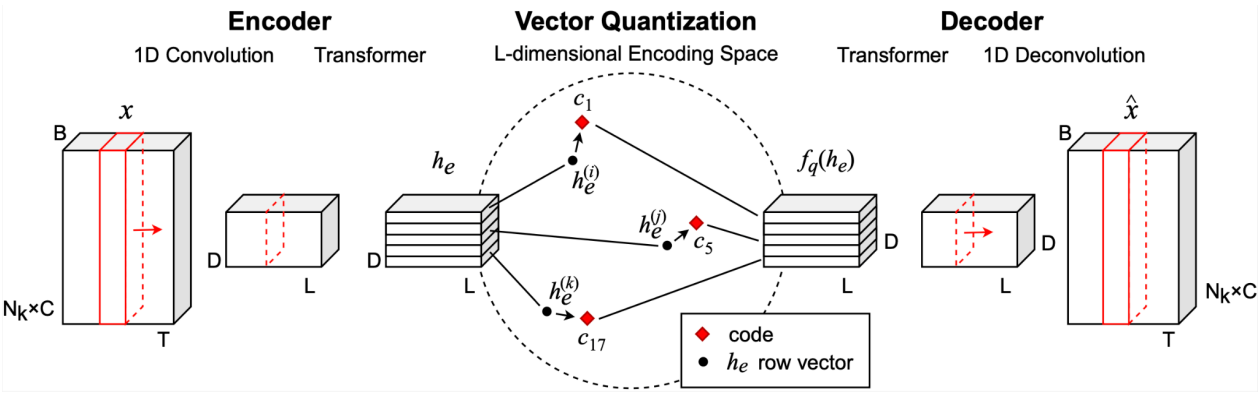

Figure 1. The modified VQ-VAE architecture (van den Oord, 2017) generates *D* vectors of shape *L* for each input, a video pose estimation (N_k, *C*, *T*). The vectors (e.g. $h_e^{(i)}$, $h_e^{(j)}$) are replaced with their nearest neighbor in a codebook (e.g. c_1, c_5) and then used to reconstruct the input.

References

Brentari, D. (1999). A prosodic model of sign language phonology. MIT Press.

Cann, R., Kempson, R., & Wedgwood, D. (2012). Representationalism and Linguistic Knowledge. In R. Kempson, T. Fernando, & N. Asher (Eds.), Philosophy of Linguistics (pp. 357-401). (Handbook of the Philosophy of Science; Vol. 14). Elsevier. http://store.elsevier.com/Philosophy-of-Linguistics/isbn-9780444517470/

Kezar, L., Pontecorvo, E., Daniels, A., Baer, C., Ferster, R., Berger, L., Thomason, J., & Sehyr, Z. S. (2023). The Sem-Lex Benchmark: Modeling ASL signs and their phonemes. In Proceedings of the 25th International ACM SIGACCESS Conference on Computers and Accessibility. https://doi.org/10.1145/3597638.3608408

Sandler, W., & Lillo-Martin, D. C. (2006). Sign language and linguistic universals: Entering the lexicon: Lexicalization, backformation, and cross-modal borrowing.

Sehyr, Z. S., Caselli, N. K., Cohen-Goldberg, A., & Emmorey, K. (2021). The ASL-LEX 2.0 Project: A Database of Lexical and Phonological Properties for 2,723 Signs in American Sign Language. The Journal of Deaf Studies and Deaf Education, 26, 263–277. https://doi.org/10.1093/deafed/enaa038

van den Oord, A., Vinyals, O., & Kavukcuoglu, K. (2017). Neural discrete representation learning. In Proceedings of the 31st International Conference on Neural Information Processing Systems (pp. 6309–6318). Curran Associates Inc.