Self-Supervised Speech Representations in a Pre-train Speech Model Represent Key Rapid Automatized Naming Variability in Autism

Sarah Ethridge[1], Joseph C.Y. Lau[1], Bronya R. Chernyak[2], Rob Voigt[1], Matthew Goldrick[1], Joseph Keshet[2], Molly Losh[1]

[1]Northwestern University, [2]Technion Israel Institute of Technology

1 Background

Individuals with autism experience significant difficulties with pragmatic language, with contributing skills often challenging to measure quantitatively with standard tools. Contributing factors to pragmatic difficulties in autism include differences in speech prosody (e.g., rate, rhythm, intonation; Patel et al., 2020), as well as differences in gaze-speech coordination that contribute to observable differences in social communication (Nayar et al., 2018). Together with differences in the phonetic properties of speech noted in autism, these factors may implicate underlying attentional and physiological differences (e.g., articulatory and visual timing) as mechanistic contributors to clinically appreciable and perceptually "odd" communication styles (e.g., reciprocity, turn-taking) in individuals with autism, their first-degree relatives, and individuals with related genetic conditions (i.e., the *FMR1* premutation; Nayar et al., 2018, 2019, 2021). Thus, fine-grained and accurate characterization of speech in autism is important for informing mechanistically focused intervention strategies grounded in a clearer etiological understanding of pragmatic differences in autism.

2 Objectives:

This study used a novel, deep-learning based measure of phonetic similarity derived from the embedding space of Hidden-unit Bidirectional Encoder Representations from Transformers (HuBERT; Hsu et al., 2021), a state-of-the-art pre-trained speech model using self-supervised learning, to represent speech differences manifested in autistic individuals relative to non-autistic controls. Variability represented through this measure was examined vis-a-vis established acoustic and performance metrics of speech and language profiles (i.e., speech rate, speech rhythm, speech errors, naming time) in autism. The ability for HuBERT to capture further variability in latent, higher order factors of autism, such as modulation of visual attention, was examined using metrics of attentional coordination of speech and gaze.

3 Methods

Analyses included speech samples from 50 autistic individuals and 45 non-autistic controls from the rapid automatized naming (RAN) task, which involved naming serial arrays of common numbers, letters, colors, and objects as quickly and accurately as possible. RAN is a deceptively simple but powerful cognitive measure that indexes speech, gaze, and their integration with important implications for pragmatic language skills in autism. Building on Chernyak et al. (2024) and Kim et al. (2025), error-free, word-sized speech samples from RAN trials were projected into the high-dimensional perceptual space of HuBERT, without the need for pre-selecting acoustic features of interest or manual alignment of speech and text samples. The distance of autistic speech samples from identical non-autistic speech samples was computed using dynamic time warping between embeddings from the 8th transformer layer of HuBERT, based on equivalent model performance across the 8-12th layers in our prior work (Chernyak et al., 2024). Using Pearson's correlations, average distance metrics were analyzed for associations with acoustic (i.e., speech rhythm and rate; Tilsen & Arvaniti, 2013), performance-based (i.e., naming time, speech error rate; Nayar et al., 2018), and gaze metrics of RAN (i.e., visual regressions, perseverations) to examine the potential link between HuBERT distance measures and the attentional coordination of speech and gaze.

4 Methods

Analyses revealed that the HuBERT distance metric was significantly correlated with the following RAN metrics: speech error rate (r (48) = 0.366, $p < 0.01$), speech rate (r (48) = -0.316, $p < 0.05$), naming time (r (48) = 0.531, $p < 0.001$), and visual regressions (r (48) = 0.424, $p < 0.01$; see

Proceedings of the Society for Computation in Linguistics (SCiL) 2025, pages 385-387.
Eugene, Oregon, July 18-20, 2025

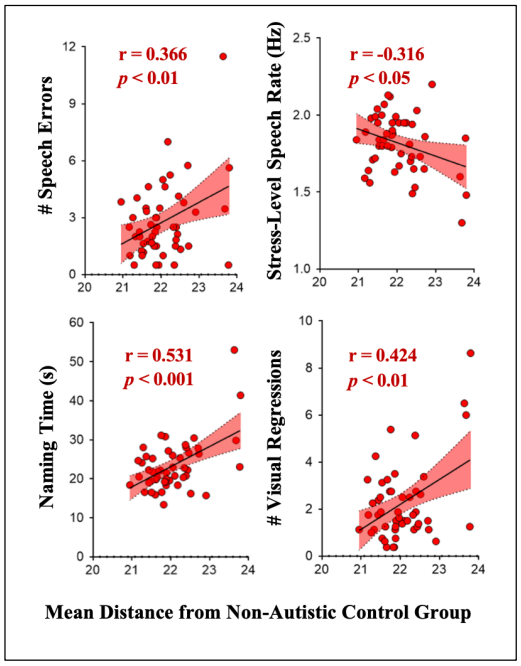

Figure 1: Associations bewteen HuBERT distance and performance, speech, and gaze metrics of rapid automatized naming in autism.

Figure 1). All significant findings survived Bonferroni correction for multiple comparisons. Variability captured by HuBERT speech representations may index subtle prosodic differences in pitch, voice quality and intensity, and articulatory variability subserving higher-order speech and language characteristics of autism, including atypical speech rhythm. Results also suggest that speech representations of HuBERT not only capture meaningful variability of speech in autism but also co-vary with eye gaze patterns that speak to the measure's sensitivity in tapping latent, higher-order linguistic and cognitive factors contributing to the communication profiles of autism.

5 Conclusions

This study demonstrates the potential utility of self-supervised pre-trained speech models, such as HuBERT, which does not require pre-defined acoustic features or speech-to-text alignment, to capture nuanced variability in the linguistic patterns of autism. The results show clear associations with meaningful variability in speech and gaze coordination, underscoring the feasibility of automating linguistic assessments in clinical settings while also providing insights into speech and its multidimensional, cross-modal relationships with broader cognitive processes in autism.

References

Chernyak, B. R., Bradlow, A. R., Keshet, J., & Goldrick, M. (2024). A perceptual similarity space for speech based on self-supervised speech representations. *The Journal of the Acoustical Society of America*, *155*(6), 3915–3929. https://doi.org/10.1121/10.0026358

Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *29*, 3451–3460. https://doi.org/10.1109/TASLP.2021.3122291

Kim, S.-E., Chernyak, B. R., Keshet, J., Goldrick, M., & Bradlow, A. R. (2025). Predicting relative intelligibility from inter-talker distances in a perceptual similarity space for speech. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-025-02652-2

Nayar, K., Gordon, P. C., Martin, G. E., Hogan, A. L., La Valle, C., McKinney, W., Lee, M., Norton, E. S., & Losh, M. (2018). Links between looking and speaking in autism and first-degree relatives: insights into the expression of genetic liability to autism. *Molecular Autism*, *9*, 51. https://doi.org/10.1186/s13229-018-0233-5

Nayar, K., Kang, X., Xing, J., Gordon, P. C., Wong, P. C. M., & Losh, M. (2021). A cross-cultural study showing deficits in gaze-language coordination during rapid automatized naming among individuals with ASD. *Scientific Reports*, *11*(1), 13401. https://doi.org/10.1038/s41598-021-91911-y

Nayar, K., McKinney, W., Hogan, A. L., Martin, G. E., La Valle, C., Sharp, K., Berry-Kravis, E., Norton, E. S., Gordon, P. C., & Losh, M. (2019). Language processing skills linked to FMR1 variation: A study of gaze-language coordination during rapid automatized naming among women with the FMR1 premutation. *PloS One*, *14*(7), e0219924. https://doi.org/10.1371/journal.pone.0219924

Patel, S. P., Nayar, K., Martin, G. E., Franich, K., Crawford, S., Diehl, J. J., & Losh, M. (2020). An Acoustic Characterization of Prosodic Differences in Autism Spectrum Disorder and

First-Degree Relatives. *Journal of Autism and Developmental Disorders*, *50*(8), 3032–3045. https://doi.org/10.1007/s10803-020-04392-9

Tilsen, S., & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: characterizing rhythmic patterns within and across languages. *The Journal of the Acoustical Society of America*, *134*(1), 628–639. https://doi.org/10.1121/1.4807565