Deep Autoencoding of Naturalistic Infant and Parent Vocalizations

Timothy M. Shea[1], Anne S. Warlaumont[1], Christopher T. Kello[1], David C. Noelle[1], Gina M. Pretzer[1], and Eric A. Walle[2]
[1]Cognitive and Information Sciences, [2]Psychological Sciences, University of California, Merced

Rationale

We view infant vocal learning as exploration for caregiver responses and ask the following:

- Can we automatically extract perceptually meaningful features from large, naturalistic audio recordings?
- **Can reconstruction error index the perceived complexity of vocalizations?**

Approach

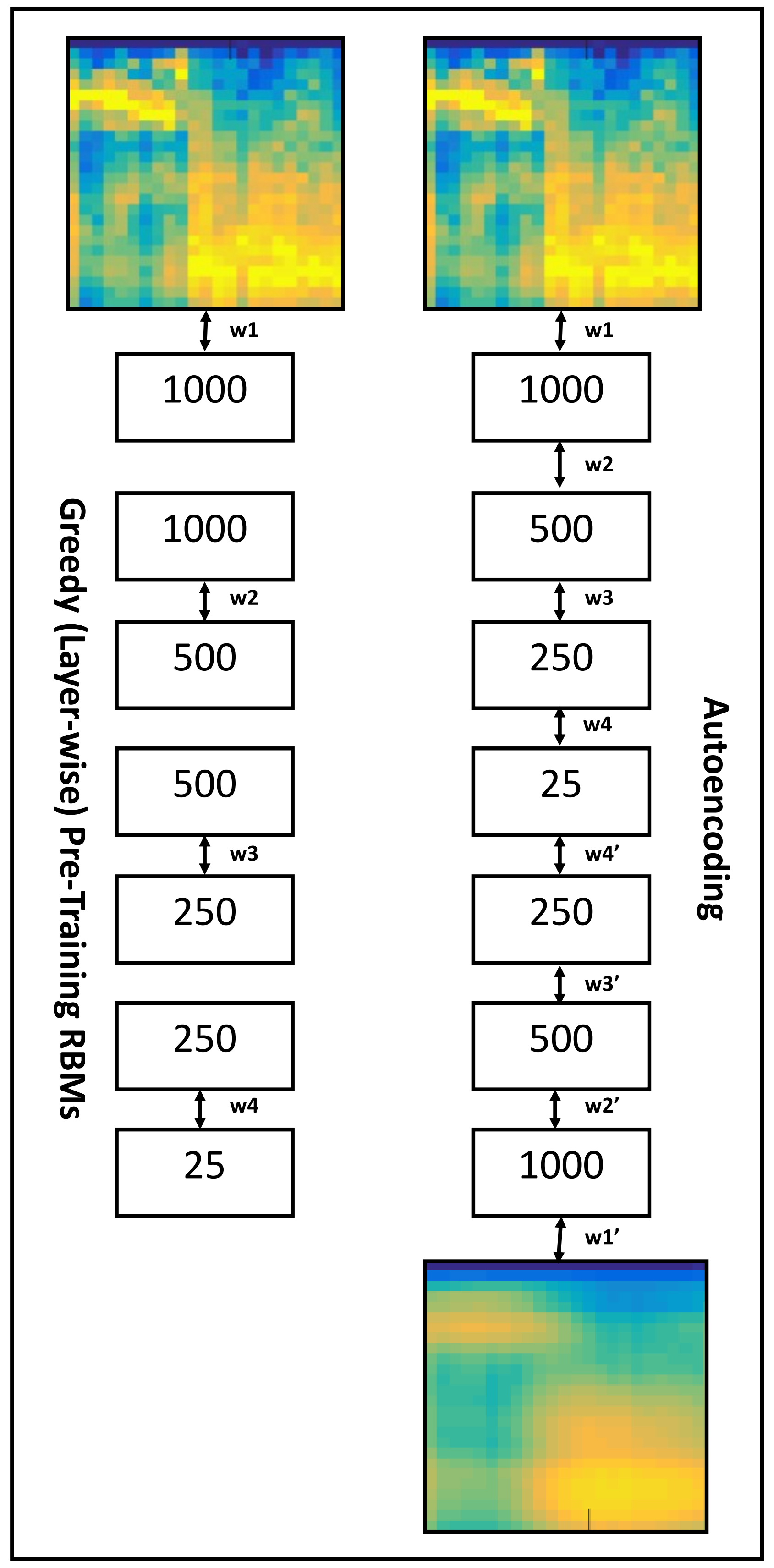
We gathered 10 day-long audio recordings from infants and caregivers. Convolutional networks and LSTMs are widely used for phoneme recognition, but not unsupervised feature extraction.
We used deep autoencoding neural networks to reduce the dimensionality of the vocalization data and analyzed the reconstructions and codes.

Conclusions

- **Reconstruction error may be a good measure of vocal complexity and similarity.**
- **Reconstruction errors suggest that vocalizations become more complex with age** and adult vocalization complexity matches infants.
- Work is ongoing to analyze hidden unit response properties and the temporal structure of encoded vocalizations.
- Future work should explore whether SRNs or LSTMs capture more relevant temporal structure.

Data Collection and Processing

10 hour audio recordings in the home, 5 infants at 3 mos. and 5 infants at 12 mos. Segmented into infant and female adult vocalizations.
Randomly selected 18000 vocalizations, truncated, Mel-scale frequency filtered in 50 ms windows, logged, normalized independently for each window.



Greedy (Layer-wise) Pre-Training RBMs
w1
1000
1000
w2
500
500
w3
250
250
w4
25

Autoencoding
w1
1000
w2
500
w3
250
w4
25
w4'
250
w3'
500
w2'
1000
w1'

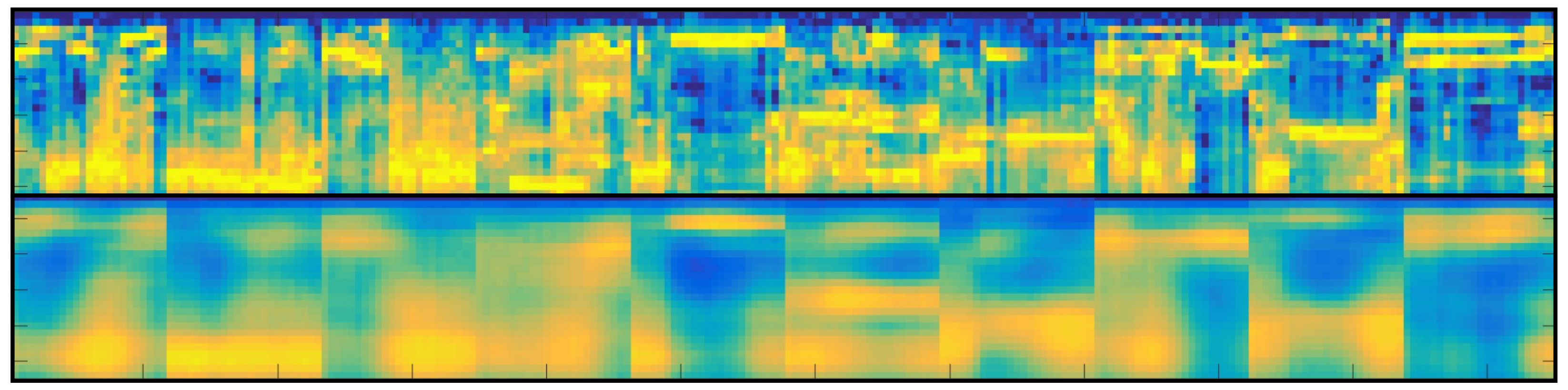
Training Deep Autoencoders

Spectrograms used to train deep autoencoders based on the method described in Hinton and Salakhutdinov[3].

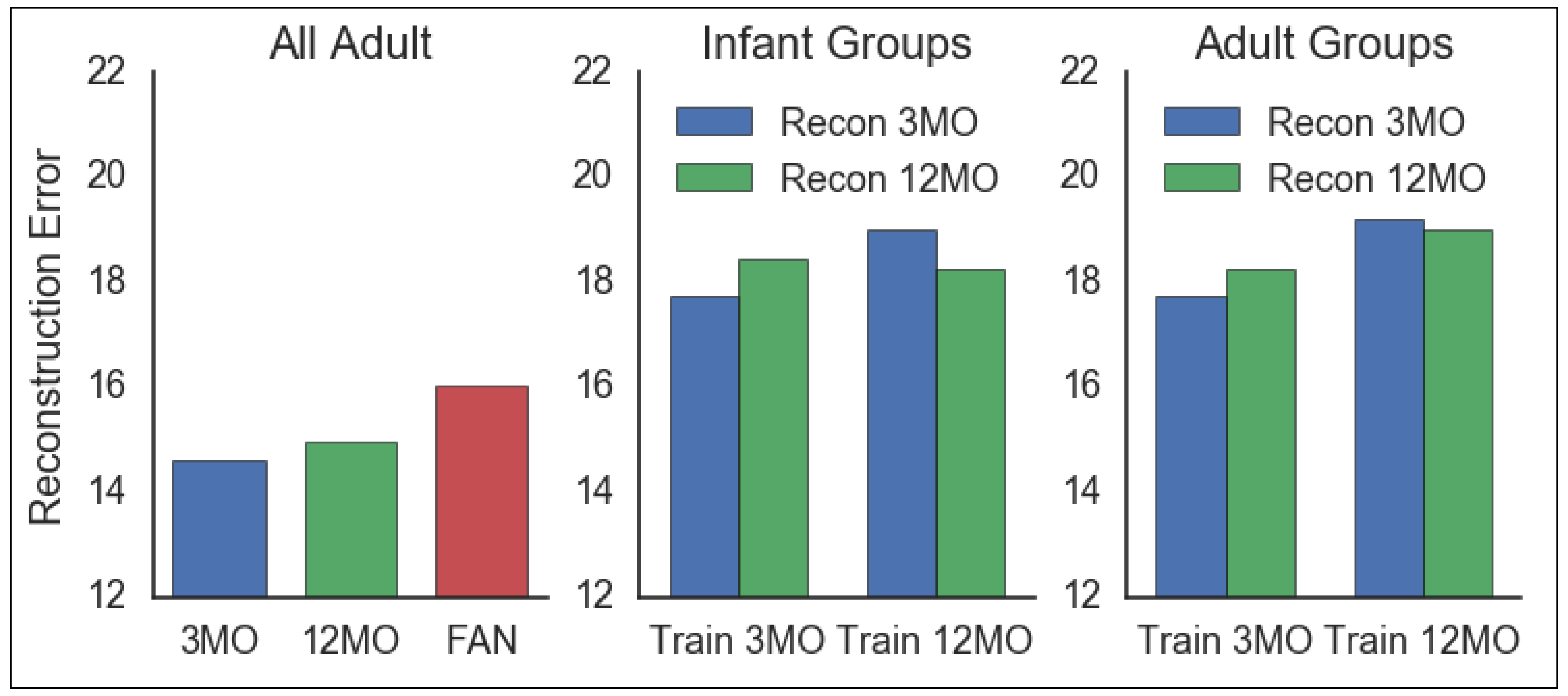
Pre-trained with contrastive divergence using Restricted Boltzmann Machines for 50 epochs.

Weights unfolded for deep autoencoder.

Trained for additional 200 epochs using conjugate gradient learning.

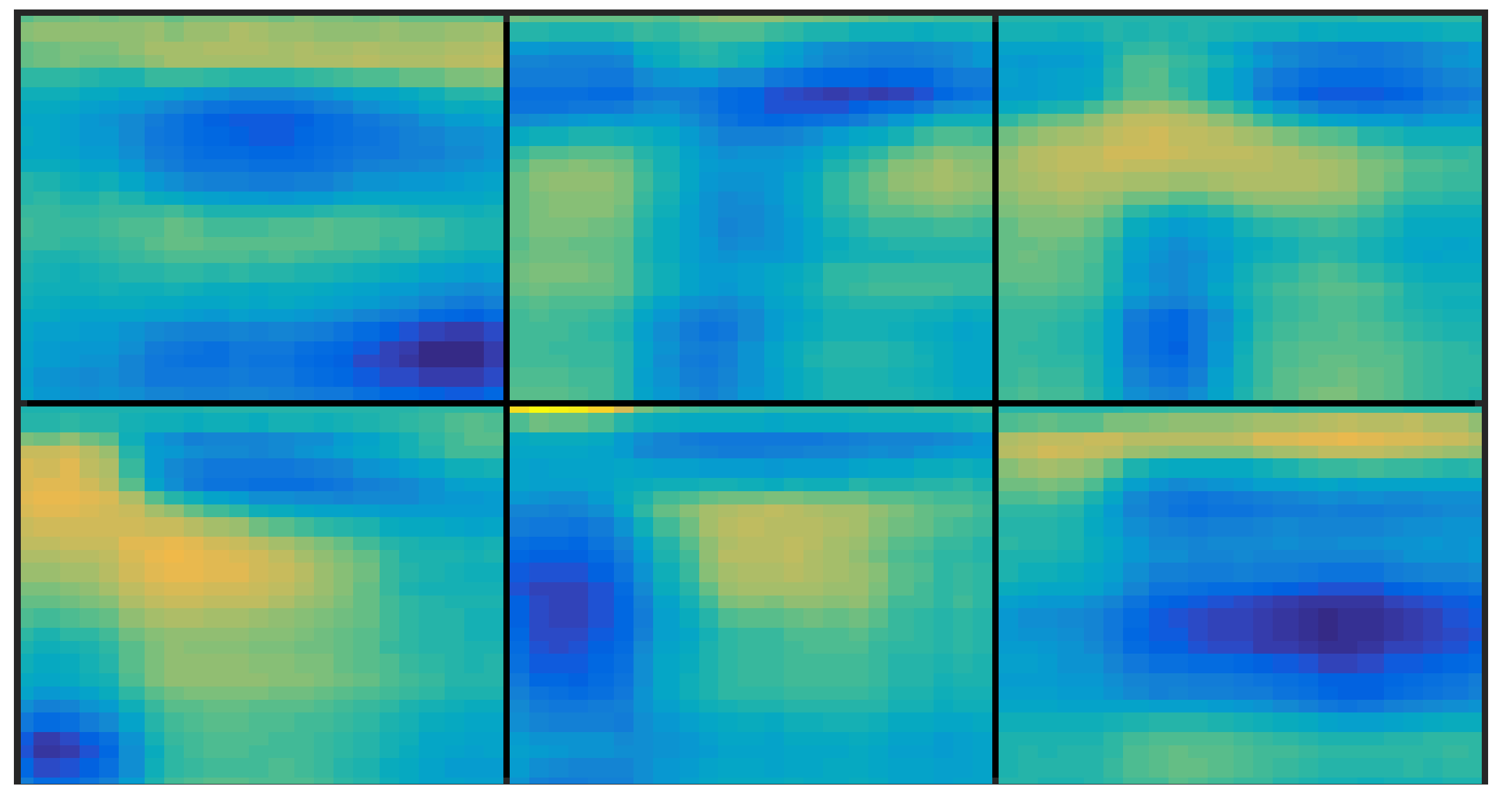


Example Spectrogram Reconstructions



Reconstruction Error Group Differences

The figure above shows reconstruction error results for five autoencoders, one trained on all adults, two trained separately on infant groups, and two trained separately on adult groups. LENA speaker categorization error potentially interferes. We are working on a validation task to eliminate this confound.



Example Hidden Unit Responses

References

Bloom, K. (1988). Quality of adult vocalizations affects the quality of infant vocalizations. Journal of child language, 15 (3), 469-480.

Golinkoff, R. M., Can, D. D., Soderstrom, M., & Hirsh-Pasek, K. (2015). (Baby) Talk to me the Social Context of Infant-Directed Speech and its effects on early language acquisition. Current Directions in Psychological Science, 24(5), 339-344.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. Science, 313 (5786), 504-507.

Warlaumont, A. S., Richards, J. A., Gilkerson, J., & Oller, D. K. (2014). A social feedback loop for speech development and its reduction in autism. Psychological Science, 25(7), 1314-1324.