Are two lexica better than one? Testing computational hypotheses with deep convolutional models

Enes Avcu[1], Olivia Newman[1], Alison Xin[2], David Gow[1,3,4]

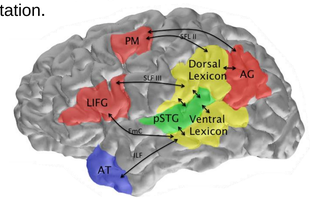
Massachusetts General Hospital/Harvard Medical School[1], Harvard University[2], Athinoula A. Martinos Center for Biomedical Imaging[3], Salem State University[4]

What We Do

Evidence from pathology, neuroimaging, behavioral paradigms provide converging evidence for the existence of parallel lexica serving the dorsal and ventral processing streams (Gow, 2012). Here, we are using deep Convolutional Neural Networks (CNNs) to examine what role computational pressures play in the emergence of this parallel architecture.

A Dual Lexicon System

- Gow's (2012) dual lexicon model synthesizes evidence from aphasia, behavioral and neural results to identify two wordform areas that mediate the mapping between acoustic-phonetic input and processing in the dorsal and ventral speech streams identified by Hickok and Poeppel (2007).
- The **dorsal lexicon,** located in the supramarginal gyrus (SMG),mediates the mapping between speech and articulation in support of speech production and the resolution of some perceptual ambiguities.
- The **ventral lexicon**, located in the posterior middle temporal gyrus (pMTG), mediates the mapping between speech and semantic/syntactic lexical representation.



A Computational Hypothesis for the Division

- Distributed feature-based lexical representations in these areas act as hidden nodes to facilitate mappings.
- We hypothesize that the complex, but systematic mapping between sound and articulation in the dorsal stream poses different computational pressures on feature sets than the more arbitrary mapping between sound and meaning.

Strategy Overview

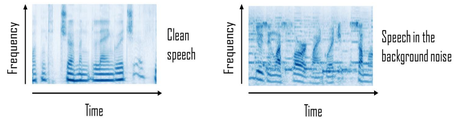
- **Identify optimal feature sets** for hypothesized mappings using CNN models explicitly trained on a large set of spoken words to recognize either wordforms or lexicosemantic representations derived from a distributional analysis of word cooccurrence (Lenci, 2018; Mandera et al. 2017).
- **Test the feature sets'** ability to pick out individual words and to support phonological versus semantic and syntactic category classification using support vector machine analyses of the feature patterns each CNN assigns to spoken word inputs.

Predictions

- CNNs trained on either dorsal (wordform) or ventral (lexical) mappings should produce features that support individual word identification because both representations yield unique features.
- Features from **CNNs trained on dorsal mappings** should have an advantage for phonological categorization but not semantic/syntactic categorization.
- Features from **CNNs trained on ventral mappings** should have an advantage for semantic/syntactic categorization but not phonological categorization.

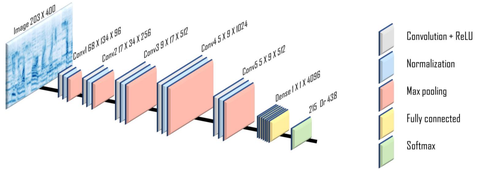
Training Data

- **Words**: 215 word were identified in the *Spoken Wikipedia Corpus* (Baumann et al., 2019)) that occurred at least 200 times and consisted of at least 4 characters.
- **Sound Files**: 2-second audio clips containing target words were extracted from the corpus. Word location within each was jittered to enlarge the training set, and each clip was combined with background noise in the form of samples of music, auditory scenes, or multi-speaker babble with moderate randomly assigned SNR levels to enhance generalization. This produced 810,000 unique sound files.
- **Cochleagrams**: Sound samples were converted into cochleagrams to simulate peripheral auditory processing and fed to the CNNs in the form of 203 x 400 cell arrays.



- **Training conditions**: Separate CNNs were trained with the same cochleagrams on different mappings:
  - **Dorsal**: Items were classified as words (215 categories)
  - **Ventral**: Items were classified based on cooccurrence with 438 collocate words identified in the billion-word *Corpus of Contemporary American English* (Davies, 2020). Each word was trained for membership in 5-19 collocate categories that each overlapped across multiple words.

Convolutional Neural Network Architecture



CNN Classification Accuracy

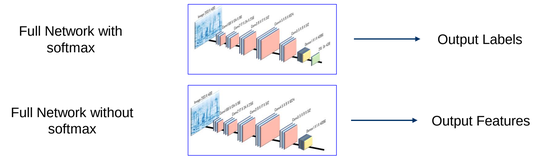
Dorsal Network (215 classes)
nearly one million images

Training ACC : 92%
Test ACC : 52%
Epoch : 100

***Ventral Network (438 classes)**
nearly six and half million images

Training ACC : 13%
Test ACC : 02%
Epoch : 14*

Feature Extraction

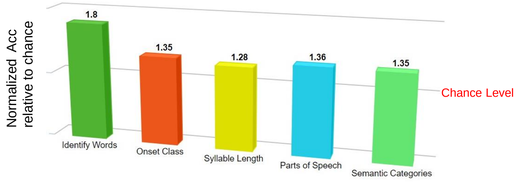


SVM Generalization Tasks

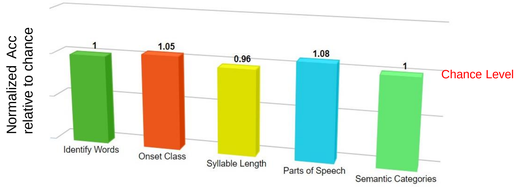
Task 1	Task 2	Task 3	Task 4	Task 5
Word Identification	Onset Class	Syllable Length	Part of Speech	Semantic Categorization
20 Classes	5 Classes	4 Classes	4 Classes	6 Classes
1000 images	5250 images	3600 images	1800 images	3300 images

SVM Discrimination

Dorsal Features



***Ventral Features (Partial Training)**



Preliminary Findings

- While the dorsal model was able to discriminate individual words with high accuracy, the ventral model needs more training.
- Featural representations extracted from the dorsal network showed classification accuracy higher than chance level.
- However, features extracted from ventral network did not show above chance level classification.

Future Steps

- The low accuracy of the ventral CNN task reflects the complexity of using a large number of overlapping phonologically heterogeneous collocate categories. Meaningful comparison of ventral and dorsal training sets will require comparable individual word identification based on unique ventral and dorsal feature encodings. This will require improvements in training and network architecture to improve validation accuracy in both CNNs.
- The underperformance of the dorsal SVM on a restricted word set relative to the performance of the dorsal CNN on word identification suggests the need for improvements in the generalization classifier.
- Following Kell et al.(2018), we ultimately hope to compare human and classifier error patterns and use the classifiers to predict cortical responses in SMG and pMTG.
- We hypothesize that feature optimization creates pressure for the emergence of multiple lexica but believe that anatomical dissociations underlying stream segregation also contribute to the emergence of dual lexica.

Bibliography

1 Gow, D. W. (2012). The cortical organization of lexical knowledge: A dual lexicon model of spoken language processing. *Brain and language, 121*(3), 273-288. doi:10.1016/j.bandl.2012.03.005
2 Hickok, G. & Poeppel (2007). The cortical organization of speech processing. Nature Reviews Neuroscience, 8, 393-402. doi: 10.1038/nrn2113
3 Lenci, A. (2018). Distributional models of word meaning. Annual review of Linguistics, 4, 151-171.
4 Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. Journal of Memory and Language, 92, 57-78.
5 Baumann, T., Köhn, A. & Hennig, F. The Spoken Wikipedia Corpus collection: Harvesting, alignment and an application to hyperlistening. Lang Resources & Evaluation 53, 303–329 (2019). https://doi.org/10.1007/s10579-017-9410-y
6 D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange and M. D. Plumbley. (2015). Detection and Classification of Audio Scenes and Events. IEEE Transactions on Multimedia 17(10), 1733-1746, 2015. http://dx.doi.org/10.1109/TMM.2015.2428998
7 Bosch, J. J., Janer, J., Fuhrmann, F., & Herrera, P. (2012). A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals in Proc. ISMIR (pp. 559-564).
8 Audiobooks, https://librivox.org
9 McDermott J. and Simoncelli E (2011). Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. Neuron (2011).
10 Feather J. and McDermott J. (2018). Auditory texture synthesis from task-optimized convolutional neural networks. Conference on Cognitive Computational Neuroscience
11 Kell, A. J., Yamins, D. L., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. Neuron, 98(3), 630-644.
12 Davies, M. (2020). *The Corpus of Contemporary American Englis*h. www.English-corpora.org/coca/.
13 Dobs, K., Kell, A., Palmer, I., Cohen, M., & Kanwisher, N. (2019). Why Are Face and Object Processing Segregated in the Human Brain? Testing Computational Hypotheses with Deep Convolutional Neural Networks. Oral presentation at Cognitive Computational Neuroscience Conference, Berlin, Germany.

Acknowledgements: This work was supported by NIDCD grant R01DC015455 and benefited from funding from NCRR grant P41RR14075. We would like to thank Alex Kell, Jenelle Feather, Ray Gonzales, Salih Tutun, Yunus Kucuk, Furkan Avcu and Arne Köhn for their invaluable advice.