Roos et al. BMC Neuroscience 2014, 15(Suppl 1):P187
http://www.biomedcentral.com/1471-2202/15/S1/P187

BMC Neuroscience

POSTER PRESENTATION Open Access

A hierarchical model of vision (HMAX) can also recognize speech

Matthew J Roos*, Michael Wolmetz, Mark A Chevillet

From The Twenty Third Annual Computational Neuroscience Meeting: CNS*2014
Québec City, Canada. 26-31 July 2014

HMAX is a well-known computational model of visual recognition in cortex consisting of just two computational operations – a "template match" and non-linear pooling – alternating in a feedforward hierarchy in which receptive fields exhibit increasing specificity and invariance [1]. Interestingly, auditory recognition problems (such as speech recognition) share similar computational requirements, and recent work in auditory neuroscience suggests that auditory and visual cortex share similar anatomical and functional organization. Based on these similarities, we tested whether HMAX could support an auditory recognition task (specifically, word spotting).

To test HMAX on word spotting, recorded speech samples from the TIMIT corpus [2] were first converted into time-frequency spectrograms using a computational model of the auditory periphery [3]. These spectrograms were then split into 750 ms frames and input to a standard HMAX model [4]. Based on observed similarities between the receptive fields in primary auditory cortex (spectro-temporal receptive fields, or STRFs) and primary visual cortex (typically modeled as oriented Gabor filters), we used S1 filters identical to those used in vision [4]. Similarly, S2 "patches" were randomly selected from C1 representations of speech sounds drawn from an independent speech corpus. One vs. all linear support vector machines (SVMs) were then trained to discriminate frames that contain a target word from those that did not. These SVMs were then tested on a novel set of test sentences using a sliding frame approach (750 ms frame size, 20 ms step size). For each frame in a sentence, the SVM produced a distance from the hyperplane, and a threshold value was applied to produce a binary classification whether or not the target word was present in the sentence. When tested on target words that appeared in a fixed context (i.e. SA sentences in TIMIT), performance was highly robust, with ROC areas consistently above 0.9. When tested on target words that appeared in variable contexts (i.e., SI sentences in TIMIT), performance was somewhat decreased with ROC areas around 0.8. This decrease in performance is likely due to the inclusion of "clutter" (i.e., target irrelevant features) within the frame, also commonly observed when HMAX is applied to visual object recognition tasks [1].

These results are novel in that they provide support for the hypothesis that the simple computational framework implemented in HMAX – consisting of a feedforward hierarchy of only two alternating computational operations – may generalize beyond vision to support auditory recognition as well. It is possible that such a representation could give rise to stable neural encodings that are invariant to behaviorally irrelevant characteristics as seen in higher order visual and auditory cortices [5,6]. While it is likely that this auditory version of the HMAX model would benefit from the use of more auditory-specific filters based on STRF models [7], the Gabor features used here are largely compatible with previous computational models based on STRFs up to the level of primary auditory cortex [8]. Additional benefit may also be gained by learning sparse representations from natural sounds, at both the S1 and S2 levels [9].

Published: 21 July 2014

References

1. Riesenhuber M, Poggio T: **Hierarchical models of object recognition in cortex.** *Nat Neurosci* 1999, **2**:1019-25.
2. Garofolo JS: **TIMIT Acoustic-Phonetic Continuous Speech Corpus.** 1993.
3. Yang X, Wang K, Shamma SA: **Auditory representations of acoustic signals.** *IEEE Trans Inf Theory* 1992, **38**:824-839.
4. Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T: **Robust object recognition with cortex-like mechanisms.** *IEEE Trans Pattern Anal Mach Intell* 2007, **29**:411-26.

* Correspondence: Matthew.Roos@jhuapl.edu
Johns Hopkins University-Applied Physics Lab, Laurel, MD 20723, USA

5. Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I: **Invariant visual representation by single neurons in the human brain.** *Nature* 2005, **435**:1102-7.
6. Chan AM, Dykstra AR, Jayaram V, Leonard MK, Travis KE, Gygi B, Baker JM, Eskandar E, Hochberg LR, Halgren E, Cash SS: **Speech-Specific Tuning of Neurons in Human Superior Temporal Gyrus.** *Cereb Cortex* 2013.
7. Theunissen FE, Sen K, Doupe AJ: **Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds.** *J Neurosci* 2000, **20**:2315-2331.
8. Mesgarani N, Shamma S, Slaney M: **Speech discrimination based on multiscale spectro-temporal modulations.** *2004 IEEE Int Conf Acoust Speech, Signal Process* 2004, **1**:601-4.
9. Hu X, Zhang J, Li J, Zhang B: **Sparsity-Regularized HMAX for Visual Recognition.** *PLoS One* 2014, **9**:e81813.

doi:10.1186/1471-2202-15-S1-P187
Cite this article as: Roos *et al.*: A hierarchical model of vision (HMAX) can also recognize speech. *BMC Neuroscience* 2014 **15**(Suppl 1):P187.