On prenatal auditory experience in humans and its relevance for machine hearing

Marin Vogelsang[1*], Lukas Vogelsang[1,2*], Sidney Diamond[1], Pawan Sinha[1]

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, United States;
[2] Brain Mind Institute, École Polytechnique Fédérale de Lausanne, Switzerland; * Equal contribution

Introduction

Given the markedly better generalization capabilities of the human perceptual system relative to computational models, a question naturally arises about the genesis of this disparity. Here, we propose that a key to robust human perception might lie in its developmental trajectory.

Unlike standard computational training procedures, perceptual development in humans appears to typically undergo a temporal progression in which sensory inputs are initially highly degraded and gain quality later on. We focus here on the auditory domain, in which this progression commences already before birth: A fetus' experience in the womb comprises low-pass filtered versions of voices and other sounds in the environment [1]. Such degraded inputs may induce the acquisition of mechanisms capable of performing extended temporal integration, facilitating robust analysis of information known to be carried by slow variations in the auditory stream, such as emotions or other prosodic content [2,3].

Computational approach

To computationally test this proposal, we assessed the consequences of training with different temporal progressions of filtered audio signals on a deep convolutional neural network's internal representations and subsequent classification of emotional prosodic content. Specifically, we trained model "M5" by [4] – a convolutional network operating directly on the raw audio waveforms – on the Toronto Emotional Speech Set [5], using four qualitatively different training regimens:

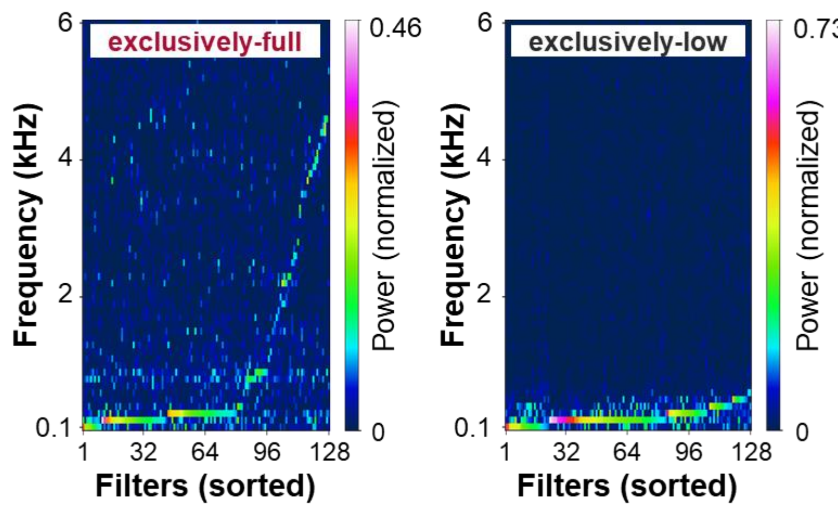
	Training epochs 0–50	Training epochs 50–100
"Exclusively-full" =	Full-frequency	Full-frequency
"Exclusively-low" =	Low-pass filtered	Low-pass filtered
"Low-to-full" =	Low-pass filtered	Full-frequency
"Full-to-low" =	Full-frequency	Low-pass filtered

Low-pass filtering was carried out at a cut-off frequency of 500 Hz, inspired by previous recordings in the womb [6].

First-layer receptive field analysis

Results of uniform training

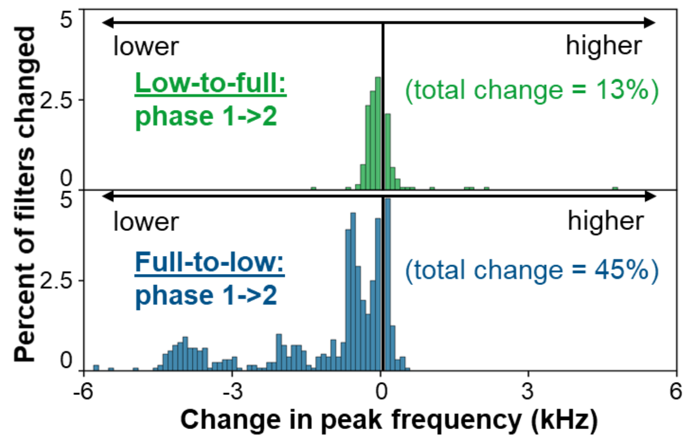
Spectral distribution of first-layer filters



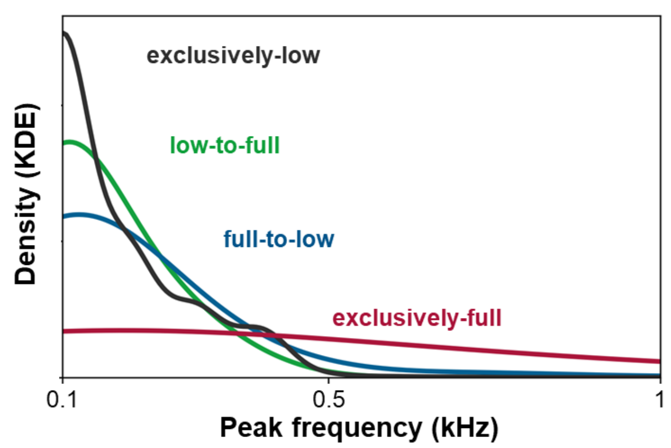
- The network trained on full-frequency inputs learned receptive fields (RFs) with a broad range of peak frequencies (left); the network trained on low-pass filtered data acquired exclusively low-frequency RFs (right).
- Training on low-frequency inputs resulted not only in more, but also purer, low-frequency filters.

Results of non-uniform training

Histogram of peak-frequency changes



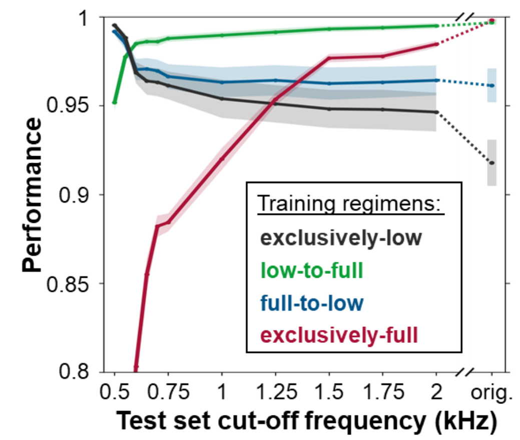
Distribution of peak frequencies



- Fewer of the RFs established during the first half of training on low-frequency inputs changed their peak frequencies upon transitioning to full-frequency inputs than when full-frequency training was later followed by low-frequency training.
- That the filters of the full-to-low network were almost exclusively enlarged temporally, in the second half of training, further resulted in the model approaching the spectral distributions of the exclusively-low and low-to-full model.

Performance-based and correlational analysis

Generalization performance



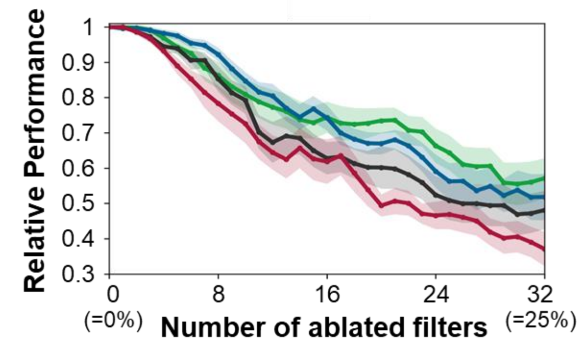
Mean and standard error of 10-fold cross-validated emotion recognition test performances

- While the model trained on full-frequency inputs exhibited poor generalization across test frequencies (red curve), the low-to-full model (green curve) yielded best generalization.
- Including low-pass filtered inputs into some phase of the training enhances performance over the exclusively full-frequency training regimen (see 'exclusively-low' and 'full-to-low' training).

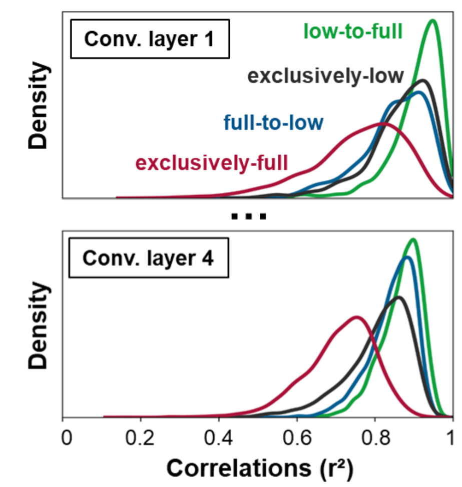
Ablation of filters with highest peak frequencies

Normalized classification performances when removing the filters with highest peak frequency

- The network trained on exclusively full-frequency inputs exhibited the strongest performance decrement upon removal of higher-frequency filters in the first layer; the low-to-full model was least affected by their removal.



Activity variation upon test frequency variation



Distribution of correlations of units' activities between low-pass filtered and full-frequency inputs

- Across all four convolutional layers, activations in the network trained on full-frequency inputs were most varied, while the low-to-full network showed the least variation, when presented with full-frequency vs. low-frequency (500 Hz) test set inputs.
- This is suggestive of stronger response stability / invariance of the low-to-full model to high-frequency modulation.

Email to: ozaki@mit.edu (Marin Vogelsang)

Conclusion / Discussion

- The simulation results suggest that the progression from low-to-full-frequency signals, rather than being an epiphenomenon, may be an enabling feature of perceptual development.
- The results also point to the utility of incorporating similar procedures into the training of computational model systems and, more generally, to the inspiration that human development may provide towards the goal of achieving more robust generalization.

References

1. Hepper, P. G., Shahidullah, B. S. (1994). Development of fetal hearing. *Archives of Disease in Childhood*, 71: F81-F87.
2. Ross, M., Duffy, R. J., Cooker, H. S., & Sargeant, R. L. (1973). Contribution of the lower audible frequencies to the recognition of emotions. *American Annals of the Deaf*, 37-42.
3. Snel, J., & Cullen, C. (2013, September). Judging emotion from low-pass filtered naturalistic emotional speech. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction* (pp. 336-342). IEEE.
4. Dai, W., Dai, C., Qu, S., Li, J., & Das, S. (2017, March). Very deep convolutional neural networks for raw waveforms. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 421-425). IEEE.
5. Dupuis, K., & Pichora-Fuller, M. K. (2010). Toronto emotional speech set (TESS). Link: https://tspace.library.utoronto.ca/handle/1807/24487
6. Gerhardt, K. J., & Abrams, R. M. (1996, February). Fetal hearing: characterization of the stimulus and response. In *Seminars in perinatology* (Vol. 20, No. 1, pp. 11-20). WB Saunders