A Model of Continuous Speech Recognition Reveals the Role of Context in Human Phoneme Perception

Gasser Elbanna[1,2], **Josh H. McDermott**[1,2,3]
{gelbanna, jhm}@mit.edu

[1]Speech and Hearing Biosciences and Technology, Harvard, Cambridge MA 02318, USA
[2]McGovern Institute for Brain Research, MIT, Cambridge, MA 02139, USA
[3]Department of Brain and Cognitive Sciences, MIT, Cambridge, MA 02139, USA

Abstract

Humans successfully transform acoustic signals into meaning despite great variability in the speech signal. The underlying mechanisms that enable such robust perception remain unclear, in part due to the absence of models that replicate human performance and that could be used to test mechanistic hypotheses. We built an artificial neural network model of continuous speech perception, optimized to recognize sequences of sub-lexical units from cochlear representations of acoustic signals. We then developed non-word recognition benchmarks to evaluate human and model speech perception. The model closely matched human performance and replicated human-like patterns of phoneme recognizability and confusions. However, human-model similarity was dependent on recurrent processing, suggesting that human recognition depends critically on bidirectional integration of information in the speech signal. The model and benchmark set the stage for future investigations into the neural and perceptual mechanisms underlying speech.

Keywords: continuous speech perception, artificial neural networks, phoneme recognition, contextual processing

Introduction

The core computational challenge of speech perception arises from the lack of a consistent one-to-one mapping between acoustic units in the signal and the sub-lexical units (e.g., phonemes, syllables) that subserve linguistic structure (Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Peterson & Barney, 1952; Perkell & Klatt, 2014). Artificial neural networks (ANNs) have helped explain human-like perceptual abilities in other domains of audition (Francl & McDermott, 2022; Saddler, Gonzalez, & McDermott, 2021), and if optimized for speech recognition might similarly yield insight into human speech perception. However, most available ANN speech models lack biological plausibility and often do not exhibit human-like patterns of performance (Weerts, Rosen, Clopath, & Goodman, 2022; Adolfi, Bowers, & Poeppel, 2023). Progress in model development is likewise hindered by the limited availability of large-scale behavioral tasks with which to compare models to humans. We sought to develop a model of continuous speech perception along with novel behavioral benchmarks to allow systematic comparisons of phoneme recognition in humans and models.

Methods

Model architecture and task objective

The model (PARROT) maps continuous speech waveforms into their constituent sub-lexical units (either characters or phonemes). The first stage is a simulation of the auditory periphery adapted from prior work (Feather, Leclerc, Madry, & McDermott, 2023), with stages of bandpass filtering, half-wave rectification, low-pass filtering, and amplitude compression to simulate auditory nerve representations. Cochlear representations are fed into six 2-dimensional convolutional layers (each with 512 channels, followed by batch normalization and ReLU) which downsample the signal to 50 Hz. The resulting representations are passed through six bi-directional Long Short-Term Memory (LSTM) layers (hidden size of 512), which capture temporal dependencies across frames. Finally, the LSTM hidden states are projected into either a 27-class character space (26 characters and a blank class) or a 40-class phoneme space (39 phonemes and a blank class) using a linear fully connected layer followed by a softmax function. The model was trained to maximize the probability of the training annotation using a Connectionist Temporal Classification (CTC) loss (Graves, Fernández, Gomez, & Schmidhuber, 2006). During inference, predicted tokens were obtained from the softmax distributions using CTC. Character/Phoneme Error Rate (C/PER) was calculated after aligning predicted and ground truth tokens using the Levenstein distance algorithm (Miller, Vandome, & McBrewster, 2009).

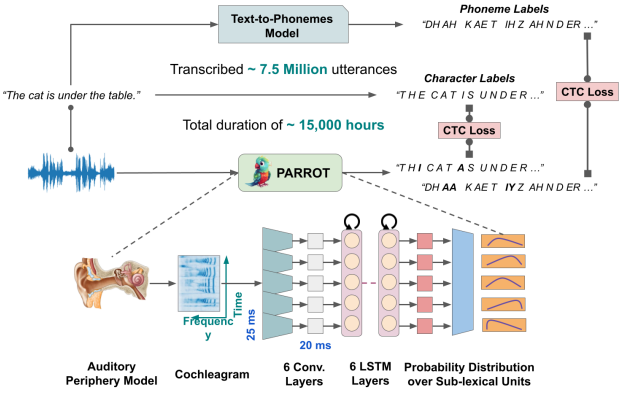

Figure 1: PARROT architecture and training pipeline.

Spoken non-word recognition experiment

To assess speech perception without linguistic influences, we designed a non-word recognition task. Participants heard synthesized non-words and typed what they heard. We obtained non-words from a pseudo-word generator that produces non-word variants from real words, abiding by English phonotactics (Keuleers & Brysbaert, 2010). We converted character transcriptions to phoneme strings using a grapheme-to-phoneme model trained on English.

Results

Models exhibit human-like phoneme recognition

We evaluated performance on the spoken non-word recognition task by computing the phoneme error rate (PER) for each non-word in the experiment. Both models performed slightly worse than humans (mean PER: PARROT(Char)=35%; PARROT(Phone)=33%; Humans=31%). Individual phonemes varied in the accuracy with which they were recognized (quantified as d'), and the phoneme-wise accuracy was highly correlated between humans and both models (PARROT(Char): r=0.97; p<0.01, PARROT(Phone): r=0.93; p<0.01; Figure 2b). The correlation remained high when calculated sepa-

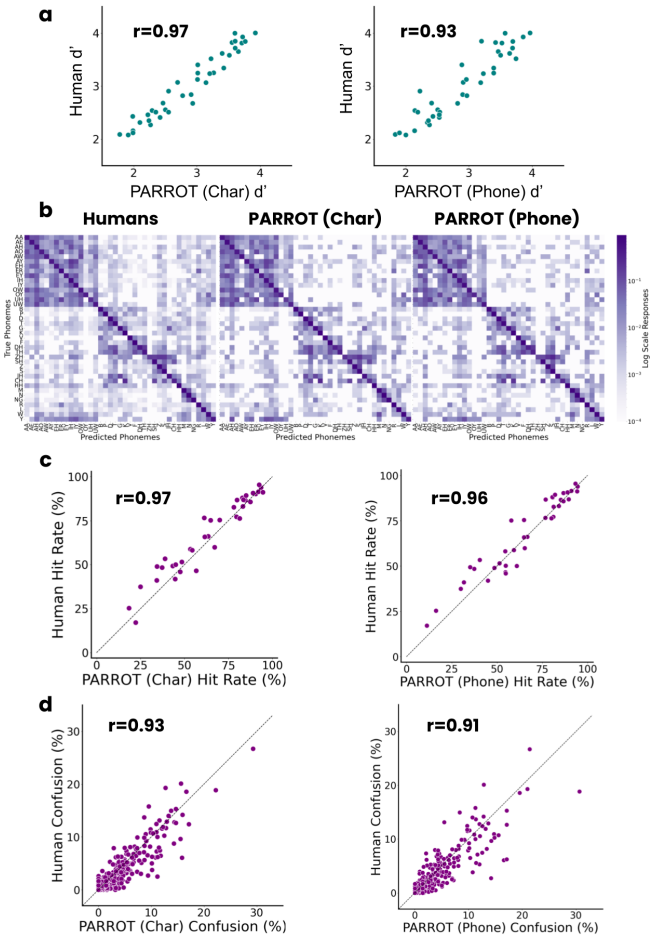

Figure 2: (a) Phoneme-wise d' for humans vs. character-based PARROT (left) and phoneme-based PARROT (right) (b) Phoneme confusion matrices in humans and models. (c-d) Phoneme-wise recognizability for humans vs. character-based PARROT (left) and phoneme-based PARROT (right), measuring hit rate, and confusion rate, respectively.

rately for consonants and vowels. The models also replicated the pattern of confusions exhibited by humans (see Figure 2a). The diagonal (hit rate) and off-diagonal entries (error patterns) of confusion matrices were each individually strongly correlated between humans and PARROT (see Figures 2c-d).

Context is critical to human-like speech perception

To investigate the mechanisms underlying the observed human-model alignment, we ablated the recurrent neural network stages and trained the model on the same task (Figure 3a). A model without recurrent-based mechanisms exhibited worse human-model alignment (Figure 3b-d).

Effect of context directionality on human-model alignment

To evaluate how the direction of contextual processing influences human-model alignment, we trained three variants of PARROT (Figure 4a): Acausal (with bidirectional LSTM layers providing access to both past and future information), Causal (unidirectional LSTM layers allowing access only to the past), and Anti-causal (unidirectional LSTM layers reversed in time,

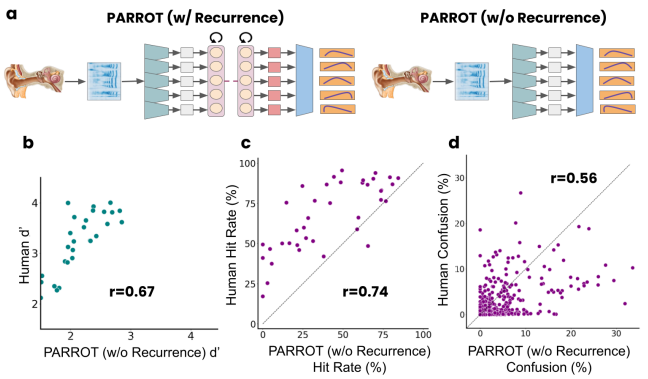

Figure 3: (a) PARROT architecture with and without recurrent layers. (b-d) Phoneme-wise recognizability between humans and PARROT without recurrence, measuring d', hit rate, and confusion rate, respectively.

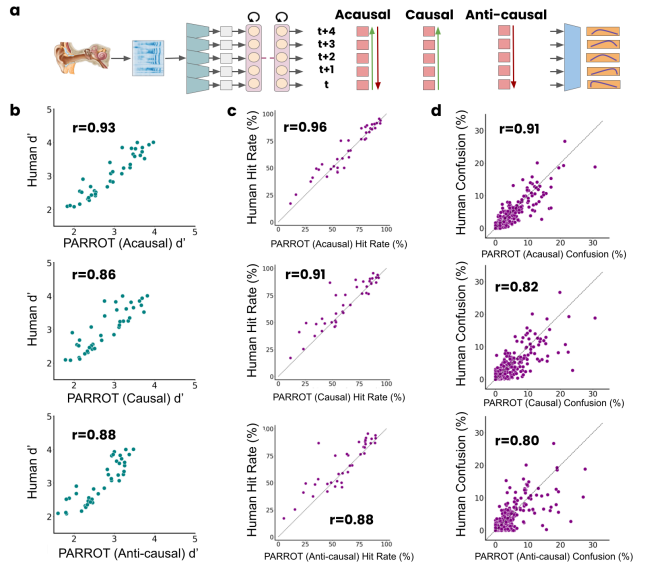

Figure 4: (a) Schematic for PARROT with different directions for contextual processing. (b-d) Phoneme-wise recognizability between humans and PARROT variants, measuring d'prime, hit rate, and confusion rate, respectively. The reported correlation coefficient is Pearson's correlation.

giving access only to future samples). The Acausal model outperformed the Causal and Anti-causal models and showed higher human-model alignment (Figures 4b–e).

Conclusions

We developed a novel deep learning model optimized to recognize sub-lexical units using a simulated cochlear front-end, that produced either character or phoneme labels. We developed new benchmarks to evaluate humans and models at a sub-lexical level, and found that the models performed similarly to humans. We then used the models to investigate contextual processing in human speech perception, finding that both past and future context was critical to obtaining human-like behavior. The results suggest that aspects of human-like speech perception emerge by optimizing for sub-lexical recognition, and that humans rely on bidirectional contextual processing to overcome the local ambiguity of speech signals.

Acknowledgment

This work was supported by National Institutes of Health (Grant number R01DC021464).

References

Adolfi, F., Bowers, J. S., & Poeppel, D. (2023). Successes and critical failures of neural networks in capturing human-like speech recognition. *Neural Networks*, *162*, 199–211.

Feather, J., Leclerc, G., Madry, A., & McDermott, J. H. (2023). Model metamers reveal divergent invariances between biological and artificial neural networks. *Nature Neuroscience*, *26*(11), 2017–2034.

Francl, A., & McDermott, J. H. (2022). Deep neural network models of sound localization reveal how perception is adapted to real-world environments. *Nature Human Behaviour*, *6*, 111–133.

Graves, A., Fernández, S., Gomez, F., & Schmidhuber, J. (2006). Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd international conference on machine learning* (pp. 369–376).

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior research methods*, *42*, 627–633.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Miller, F. P., Vandome, A. F., & McBrewster, J. (2009). *Levenshtein distance: Information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance.* Alpha Press.

Perkell, J. S., & Klatt, D. H. (2014). *Invariance and variability in speech processes*. Psychology Press.

Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *The Journal of the acoustical society of America*, *24*(2), 175–184.

Saddler, M. R., Gonzalez, R., & McDermott, J. H. (2021). Deep neural network models reveal interplay of peripheral coding and stimulus statistics in pitch perception. *Nature Communications*, *12*, 7278.

Weerts, L., Rosen, S., Clopath, C., & Goodman, D. F. (2022). The psychometrics of automatic speech recognition. *bioRxiv preprint bioRxiv:2021.04.19.440438*.