A Model of Speech Recognition Reproduces Signatures of Human Speech Perception and Reveals Novel Mechanisms

Gasser Elbanna, Arnav Aggarwal, Josh McDermott

Background: Humans rapidly and dexterously transform acoustic waveforms into linguistic representations despite the inherent variability of the speech signal. The mechanisms that enable such robustness remain poorly understood, in part due to the lack of stimulus-computable models that both approach human performance and can test theoretical principles and mechanistic hypotheses.

Methods: We built an artificial neural network model of continuous speech recognition optimized to produce sequences of sublexical units (phonemes or characters) from simulated cochlear representations. The front end simulated outer, middle, and inner ear filtering, followed by half-wave rectification and low-pass filtering at 4 kHz to approximate the upper limit of mammalian auditory-nerve phase-locking, yielding cochleagrams that served as input to stacked convolutional and recurrent layers trained to predict sublexical sequences. The model was trained on 15,000 hours of speech data superimposed on naturalistic noisy backgrounds, with varied SNRs and sound levels.

We compared the model to humans using a new large-scale benchmark of nonword transcription and discrimination along with a battery of previously documented signatures of speech perception, including categorical perception, neighbourhood density effects, phonotactic probability effects, formant-based vowel space, and benefits of talker familiarity. We tested the role and importance of contextual integration by ablating contextual integration mechanisms (recurrent processing), and by manipulating the directionality of contextual processing in the model.

Results: Human nonword transcription and discrimination exhibited highly reliable phoneme confusion structure (split-half r=0.98). The model closely matched human performance, reproducing human-like patterns of phoneme recognizability (r=0.96) and confusions (r=0.95). It also recapitulated classic human perceptual signatures: categorical perception of consonants, sensitivity to neighborhood density and corpus statistics, encoding F1-F2 vowel space at earlier stages of the model and improved recognition with talker consistency.

Removing contextual processing substantially reduced human–model similarity, suggesting that human recognition depends critically on integrating surrounding speech information.

Conclusions: We built a biologically grounded stimulus-computable model of speech perception, and developed a suite of evaluations with which to compare it to humans and probe the basis of recognition. The model aligns closely with human behavior on speech perception tasks, but only if allowed to integrate information from the surrounding speech signal, implicating contextual processing in human speech perception. By providing both a working model and an evaluation suite, this work supports systematic assessment of existing and future theories of speech perception.