Parallels and divergences in spoken word recognition between humans and a neural network model.

Máté Aller and Matthew H. Davis MRC Cognition and Brain Sciences Unit, University of Cambridge, Cambridge , UK

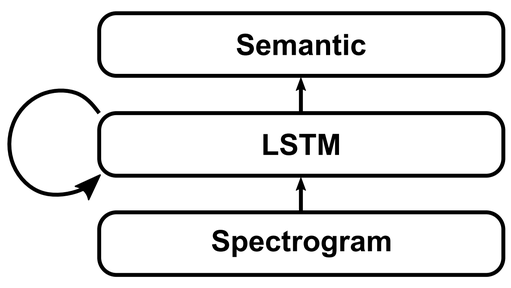
Introduction

Recent advances in artificial neural networks (ANNs) enabled new models of human speech perception which can operate on real speech (as opposed to previous models which mainly used symbolic or simplified speech representations, e.g. phonetic features).

We compared how an ANN model of human speech perception responds to changes in speech rate and speakers, which are well studied phenomena in human speech perception research.

As a model of human speech perception, we used a recently published end-to-end model with a single recurrent hidden layer of long short-term memory (LSTM) nodes (coined ‘EARSHOT’, (Magnuson et al., 2020))

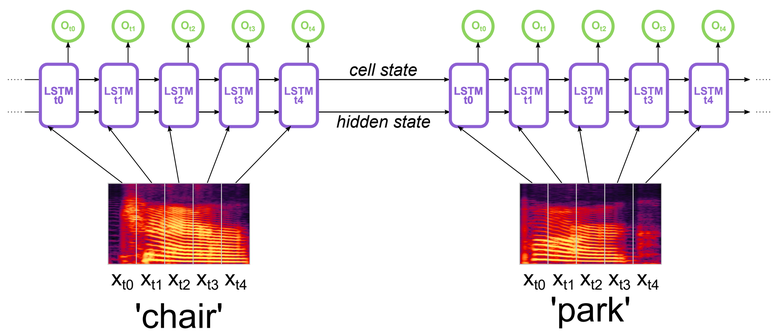
EARSHOT



We aim to identify model design features critical for human-like behavioural performance and thereby better understand the human speech perception architecture.

Model improvements

Model memory spans across words:
LSTM hidden and cell states are propageted across words as well as time steps. In this way EARSHOT can potentially capture how prior context influences word recognition (i.e., speaker identity or speech rate).



Better control for overfitting:
Using regularization (dropout) in the recurrent (LSTM) layer.

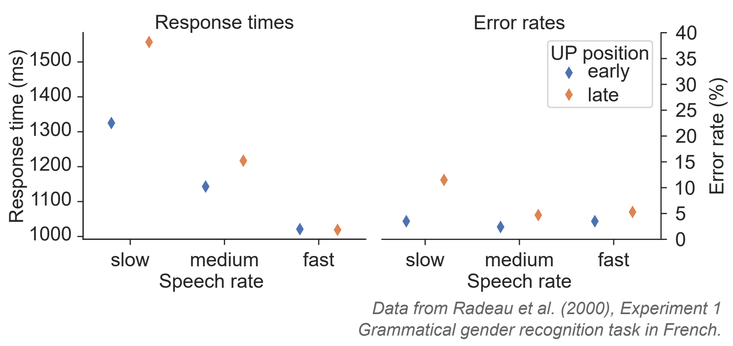
Changed input features:
- Logarithmically spaced spectral channels (Mel) instead of linearly spaced.
- Less input features (64 vs 256)

Extended training data:
- 1161 words produced by 15 speakers from Apple text-to-speech.
- Augmented by including slow (x0.5) and fast (x2.0) produced version of each word.

Uniqueness point effects

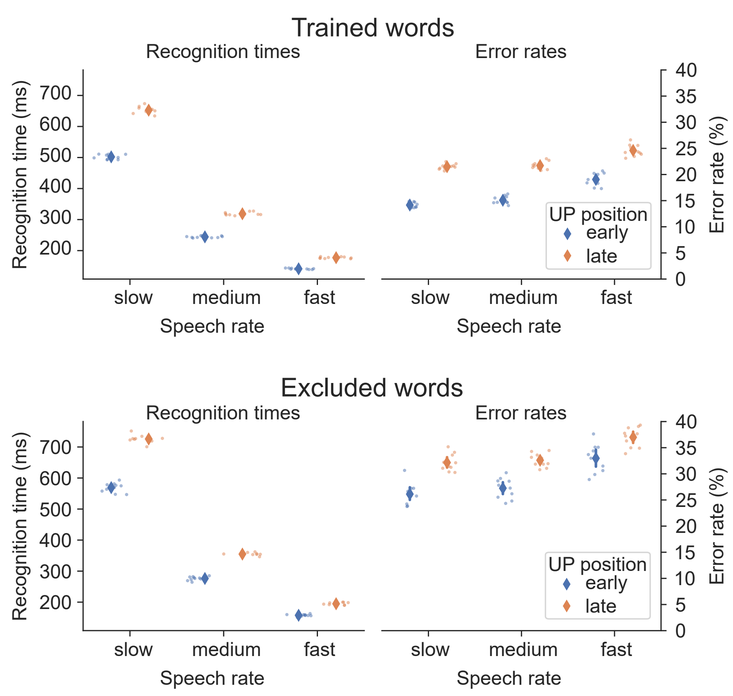
Human listeners recognize words with early uniqueness points (UP) earlier than those with later UPs. This effect crucially depends on speech rate such that for slow speech the advantage for early UP words is increased and it disappears at high speech rate (Radeau et al. (2000)).

Human listeners



Data from Radeau et al. (2000), Experiment 1
Grammatical gender recognition task in French.

EARSHOT



Similarly to humans, EARSHOT also shows:
- Recognition time (RT) advantage on early compared to late UP words (main effect of UP location: F(1,9)=5500.53, p<.0001).
- This depends on speech rate (interaction between UP location and speech rate: F(2,18)=2768.96, p<.0001).

Unlike humans, EARSHOT shows:
- RT advantage even at fast speech rate (paired t-test early vs. late UP locations for fast speech rate: t(9)=36.95, p<.0001).
- Much higher error rates, although error rate patterns are similar to humans.

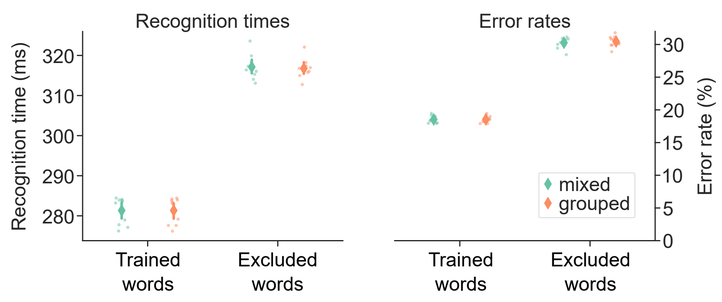
Speaker change effects

Human listeners show a consistent cost in response times when different speakers are mixed compared to when they are grouped (Magnuson et al. (2021))

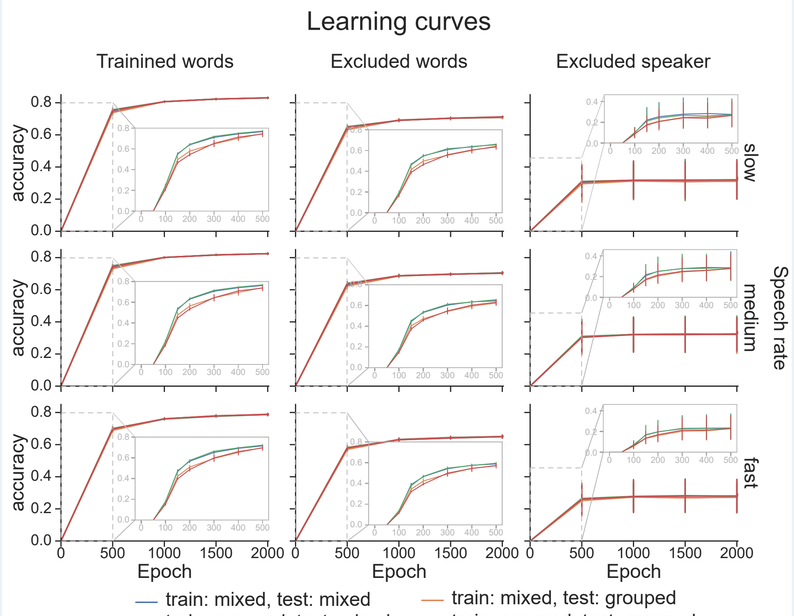
Human listeners

Data from Magnuson et al. (2021), Experiment 1
Target monitoring task on morae (syllables) in Japanese.

EARSHOT



Unlike humans, EARSHOT, when fully trained, does not show an RT cost of speaker change:
- Recognition times (main effect of grouping: F(1,9)=1.09, p=.3247)
- Error rates (F(1,9)=1.18, p=.3050)



During training (100-500 epochs) EARSHOT shows worse performance when words are grouped by speaker compared to when mixed (main effect of grouping: (F(1,5)=73.42, p=.0004). This difference disappeares when performance reaches plateau (after about 500 epochs).

Discussion and planned work

In summary, EARSHOT responds similarly to humans to speech rate changes, but not to speaker changes.

The lack of RT advantage for fast speech in humans perhaps reflects experimental limitations, we are planning follow-up experiments to clarify this.

We are investigating which architectural features could explain the lack of speaker change effects.

We are planning further simulation experiments to investigate how EARSHOT responds to degraded speech since this also influences uniqueness point effects in humans.

References:

Magnuson, J. S., Nusbaum, H. C., Akahane-Yamada, R., & Saltzman, D. (2021). Talker familiarity and the accommodation of talker variability. Attention, Perception, & Psychophysics, 83(4), 1842–1860. https://doi.org/10.3758/s13414-020-02203-y

Magnuson, J. S., You, H., Luthra, S., Li, M., Nam, H., Escabí, M., Brown, K., Allopenna, P. D., Theodore, R. M., Monto, N., & Rueckl, J. G. (2020). EARSHOT: A Minimal Neural Network Model of Incremental Human Speech Recognition. Cognitive Science, 44(4), e12823. https://doi.org/10.1111/cogs.12823 https://github.com/maglab-uconn/EARShot_TF2

Radeau, M., Morais, J., Mousty, P., & Bertelson, P. (2000). The Effect of Speaking Rate on the Role of the Uniqueness Point in Spoken Word Recognition. Journal of Memory and Language, 42(3), 406–422. https://doi.org/10.1006/jmla.1999.2682