Do neural speech models show human-like linguistic biases in speech perception?

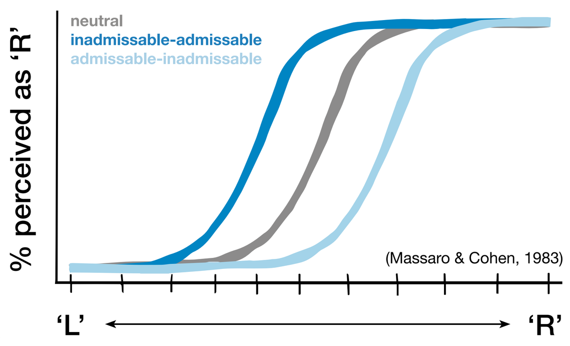
Marianne de Heer Kloots*[,1], Willem Zuidema[1]

[1]Institute for Logic, Language and Computation; University of Amsterdam, The Netherlands
*m.l.s.deheerkloots@uva.nl

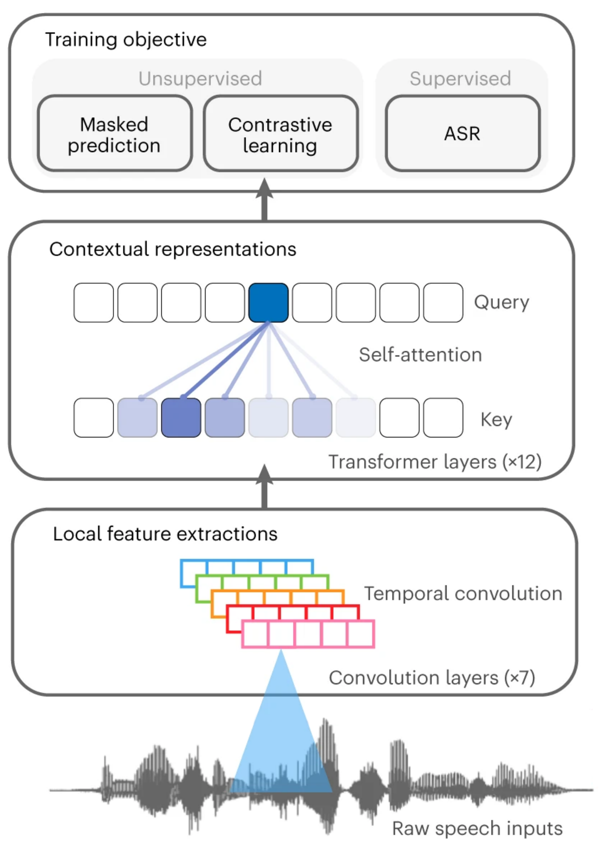
Human speech sound categorization is *linguistically informed*

For example by **phonotactic admissability**:
In English, *TL vs. TR
SL vs. *SR

When hearing acoustically ambiguous speech sounds, humans are biased towards perceiving the most likely phoneme given the surrounding phonotactic context[1].



Neural speech models like Wav2Vec2[2] operate on the raw waveform and are pre-trained on a *self-supervised* masked audio segment prediction task.



➡ Do similar perceptual biases emerge in Wav2Vec2?

And how can we localize them?

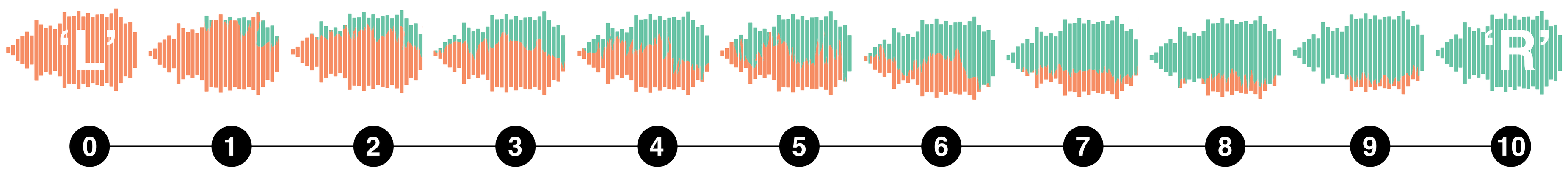
We compare 7 Wav2Vec2 models

4 base models (12 layers):
- untrained
- pre-trained on acoustic scenes
- pre-trained on speech
- pre-trained on speech & fine-tuned on text transcription

3 large models (24 layers)
- untrained
- pre-trained on speech
- pre-trained on speech & fine-tuned on text transcription

Using a controlled set of stimuli

- 11-step acoustic continua between /l/ and /r/



- interpolating on fundamental frequency, spectral envelope, and aperiodic component parameters with the WORLD vocoder GUI[3]
- 3 phonotactic contexts:

/l/ inadmissable
/r/ admissable

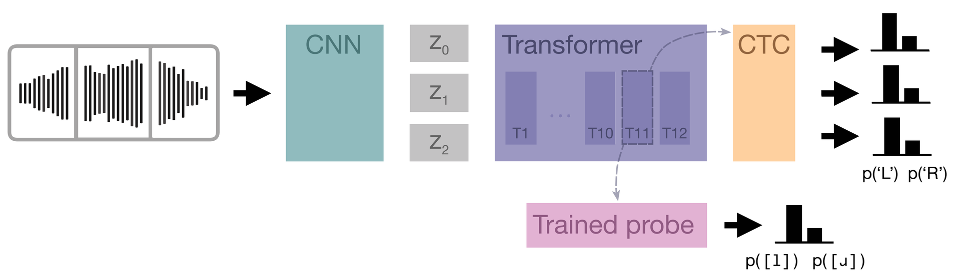
/r/ inadmissable
/l/ admissable

both unlikely

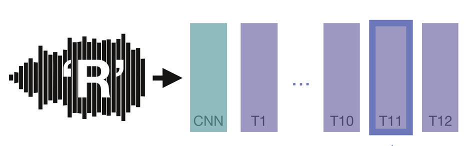
- 2 voices (Google TTS `en-US-Standard-A` and `en-US-Standard-E`)

And 3 analysis methods

- **Probing classifier probabilities**
  Binary logistic regression probes trained on 4000 phonetically transcribed word pronunciations from TIMIT

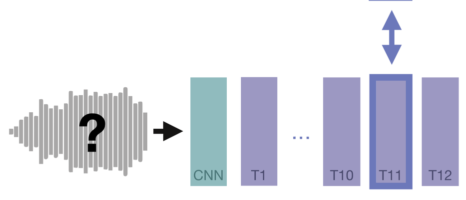


- **CTC-lens probabilities**
  Output of the text-transcribing CTC head when processing the hidden states from intermediate Transformer blocks



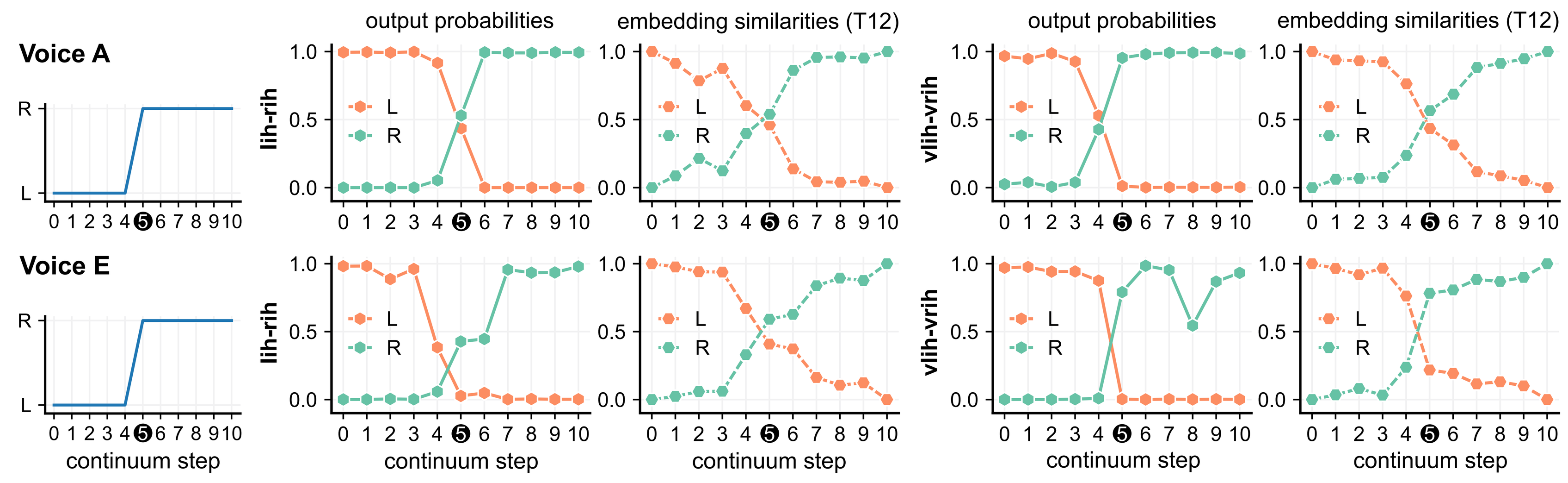
- **Embedding similarities**
  Based on cosine distances between hidden states for the morphing target sound (X) and the unambiguous continuum endpoints

$$sim(X, \text{'R'}) = 1 - \frac{D_{cos}(X, \text{'R'})}{D_{cos}(X, \text{'R'}) + D_{cos}(X, \text{'L'})}$$

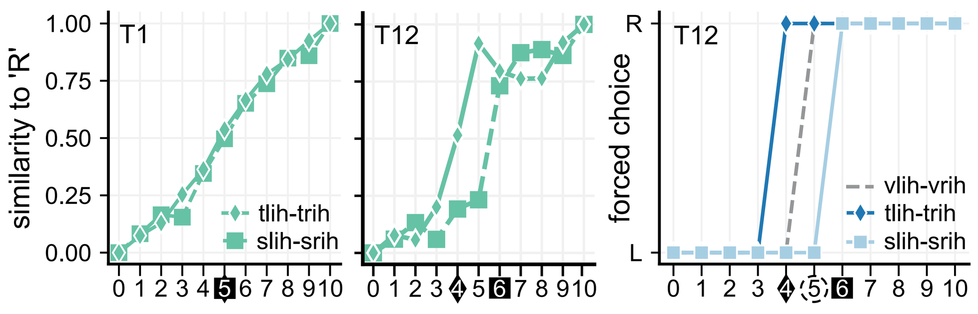


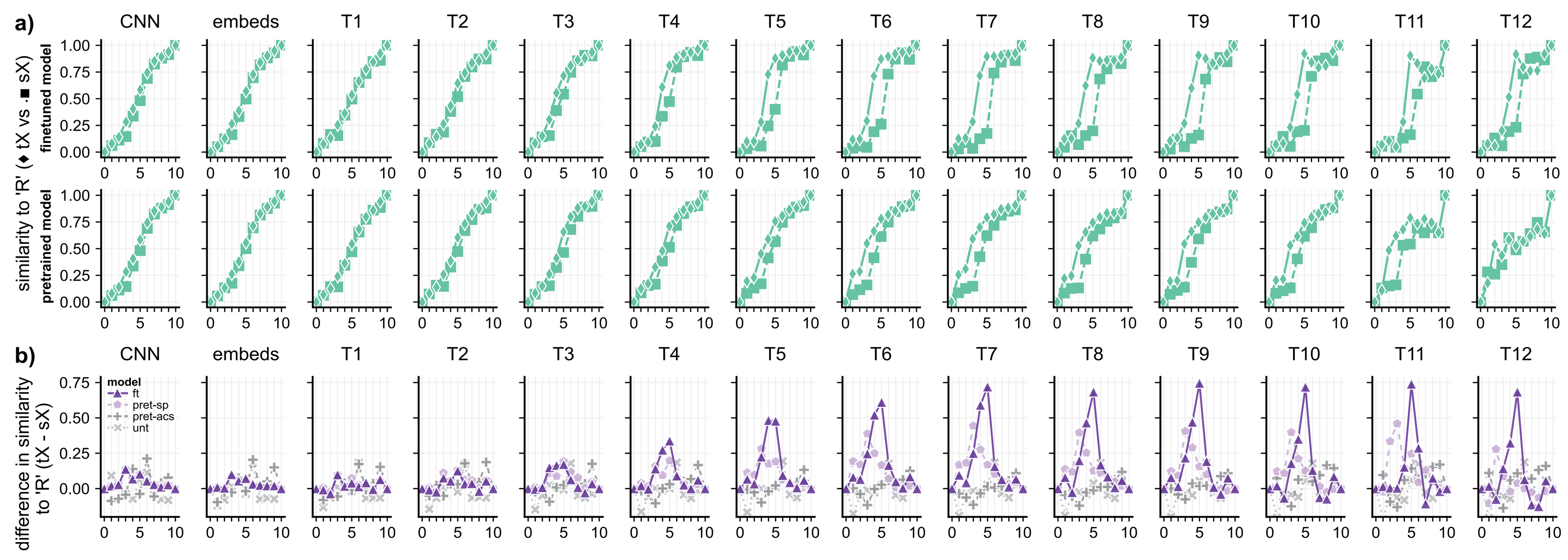
Results

1. In the ASR-finetuned model, character output probabilities are aligned with final layer embedding similarities



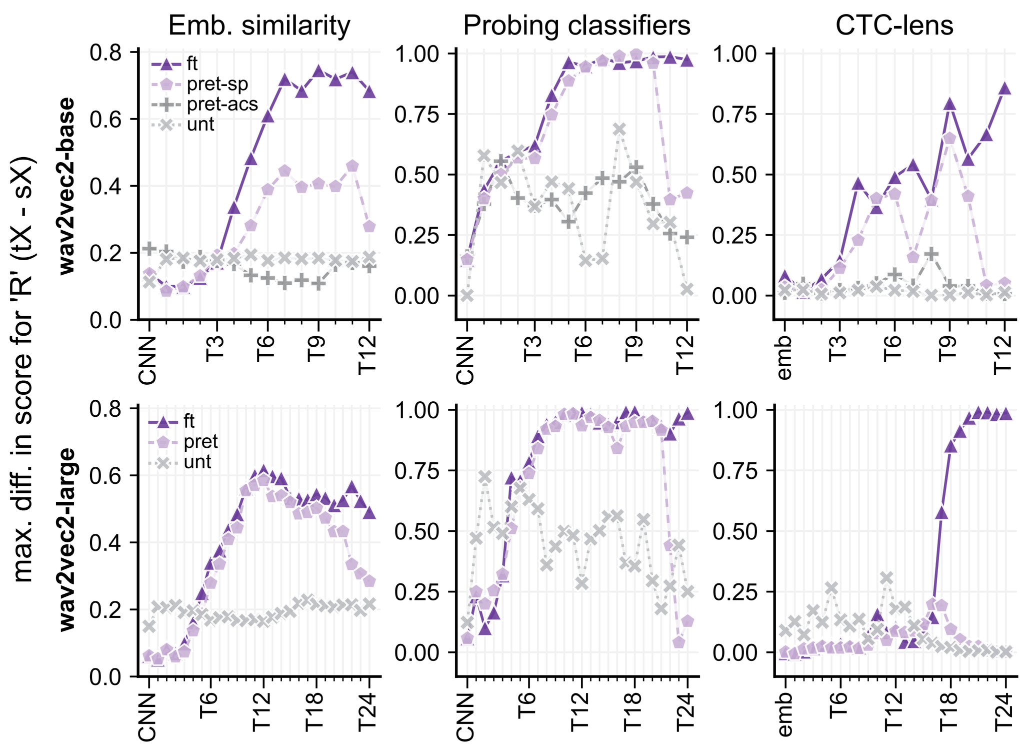
2. Sensitivity to phonotactic context emerges around layer 4 of the model's Transformer module





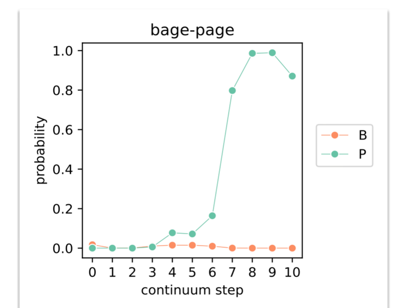
3. Comparing models and analysis methods:

- Phonotactic sensitivity is amplified by ASR finetuning, but also present in fully self-supervised models when pre-trained on speech (but not acoustic scenes)
- The embedding similarity measure is most sensitive to distinct characteristics of different models' representational spaces
- The CTC-lens measure deviates from the other analysis measures in the large model architecture — phonological information encoded in earlier layers may only later get transformed into a format that the CTC head can map to orthographic predictions



Conclusions & Next steps

- Internal representations of Wav2Vec2 models trained on English speech show human-like adaptation to phonotactic constraints
- A **symbolic training objective like character prediction is not necessary** for the Wav2Vec2 model to implicitly learn information about English phonotactic structure
- Similar phonetic categorization paradigms will allow us to examine the presence of more abstract (e.g., **lexical** and **syntactic**) biases, and their robustness across different model architectures



References

[1] Massaro, D. W., & Cohen, M. M. (1983). Phonological context in speech perception. *Perception & psychophysics, 34*(4), 338-348. https://doi.org/10.3758/BF03204046
[2] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems, 33*, 12449-12460. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html
[3] Kawahara, H., & Morise, M. (2024). Interactive tools for making vocoder-based signal processing accessible: Flexible manipulation of speech attributes for explorational research and education. *Acoustical Science and Technology, 45*(1), 48-51. https://doi.org/10.1250/ast.e23.52
[4] Garofolo, John S., et al. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1. *Philadelphia: Linguistic Data Consortium.* https://catalog.ldc.upenn.edu/LDC93S1