Investigating language learning trajectories in a self-supervised speech model

Marianne de Heer Kloots[1]*, Martijn Bentum[2], Hosein Mohebbi[3], Charlotte Pouw[1], Gaofei Shen[3], Willem Zuidema[1]

[1]Institute for Logic, Language, & Computation; University of Amsterdam, The Netherlands
[2]Centre for Language Studies; Radboud University, The Netherlands
[3]Department of Cognitive Science and Artificial Intelligence; Tilburg University, The Netherlands

*m.l.s.deheerkloots@uva.nl

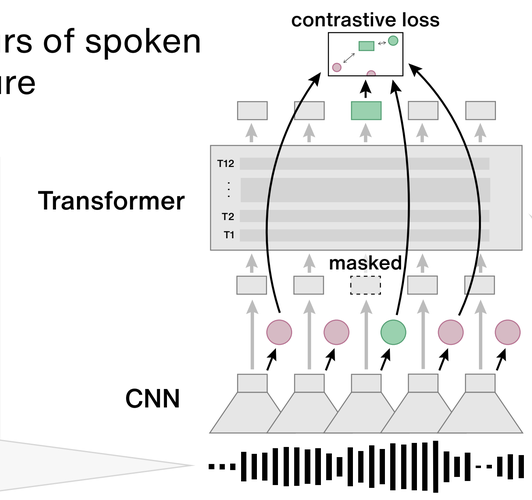
Where & when do different levels of linguistic structure get encoded in models learning from speech?

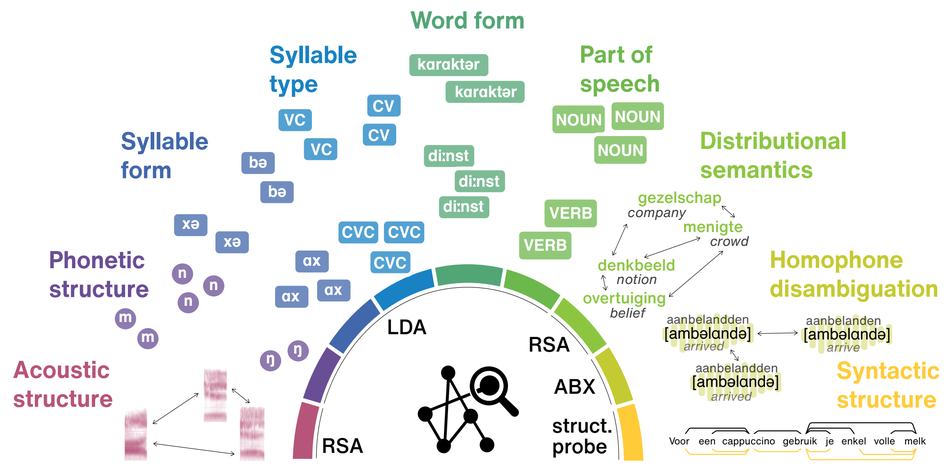
We train a new Wav2Vec2-base model on 960 hours of spoken Dutch, and probe it for 9 levels of linguistic structure

Corpus Gesproken Nederlands (Spoken Dutch Corpus) 537 hours

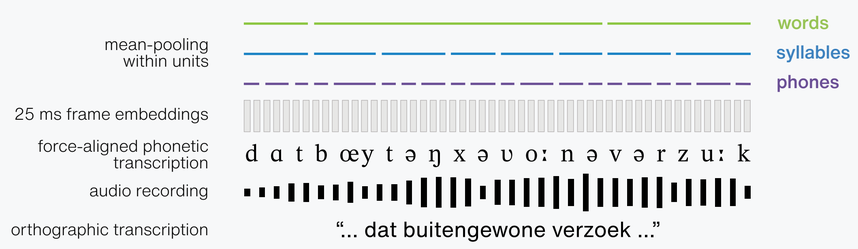
MultiLingual Librispeech 211 hours

CommonVoice 212 hours

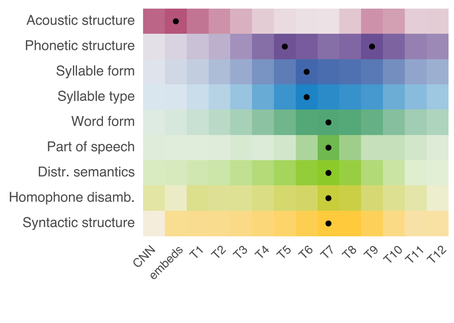


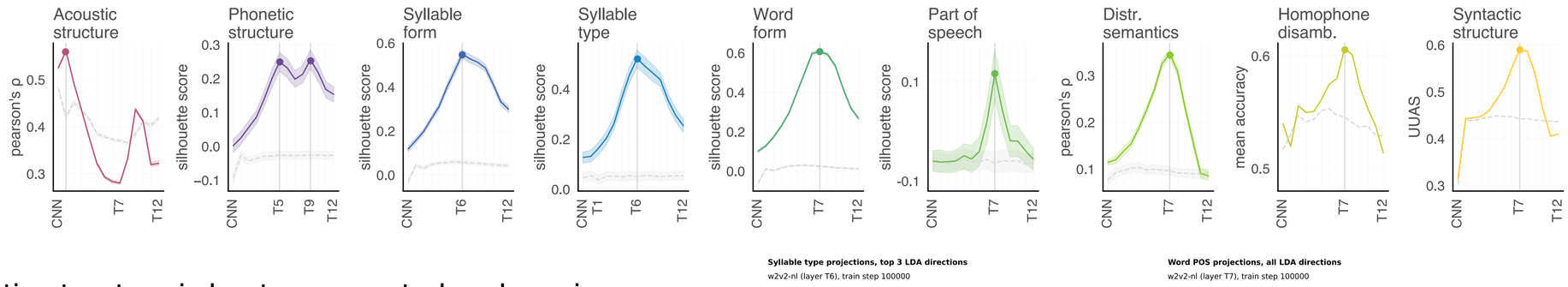


Embedding extraction:



Different levels of linguistic structure... ... are distributed across model layers



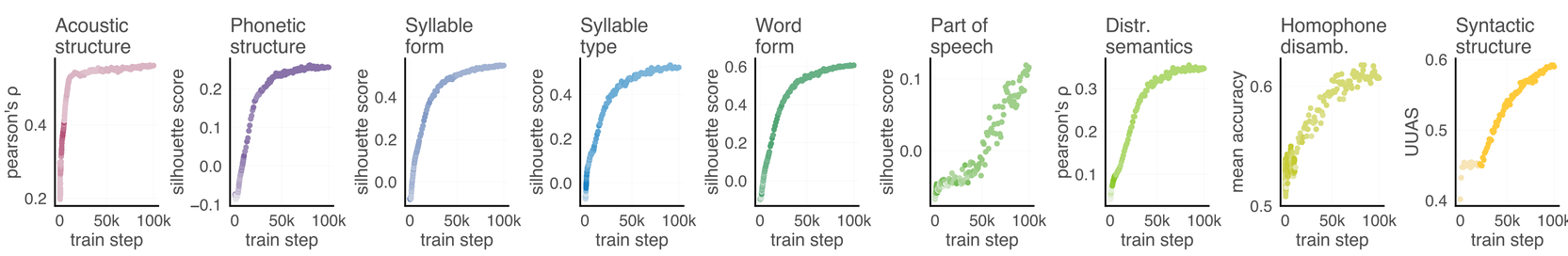


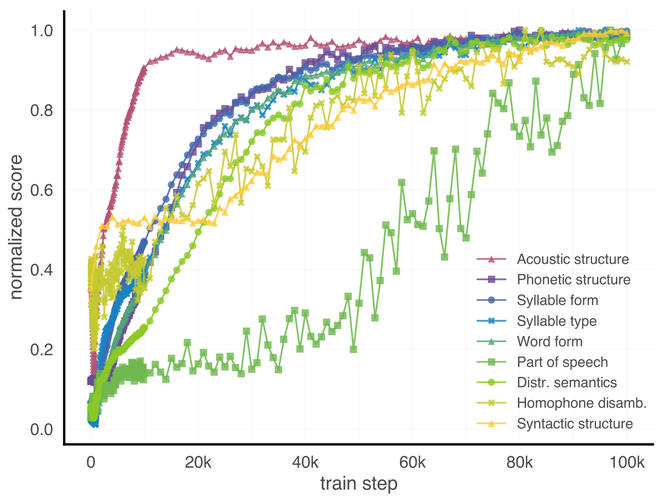
After 100k training steps, acoustic structure is best represented early on in the model's processing hierarchy, followed by phonetic and syllabic structure. Lexical and syntactic structure are jointly concentrated in a later model layer. Interestingly, a second phonetic structure peak occurs after the lexical and syntactic peaks but before the final model layer.

Linear Discriminant Analysis allows us to inspect the discriminant spaces for each fitted projection

... and show distinct learning trajectories during model training

Early stages of model training are characterized by learning generally useful representations of speech acoustics. Phonetic, syllabic, word form, and distributional semantic structure show a gradual increase from the start of model training. The capacities to disambiguate homophones, to represent part of speech classes and to identify syntactic dependencies only start developing in later training stages.

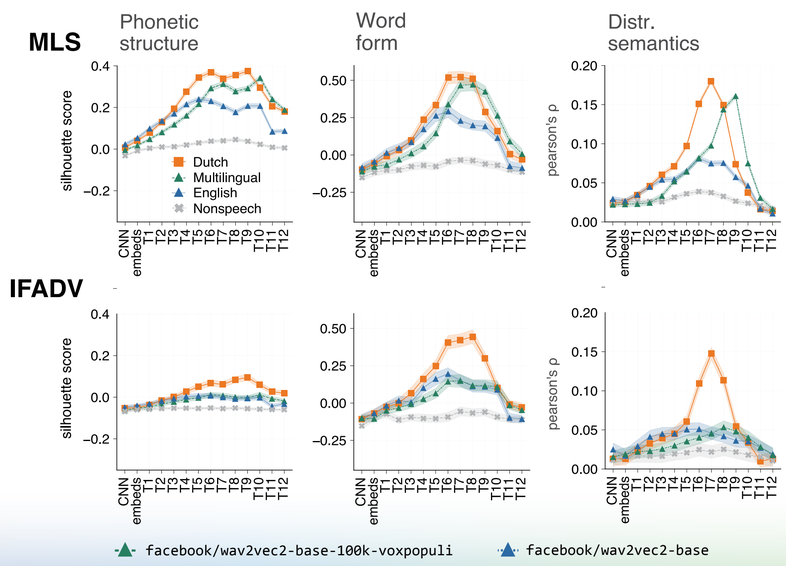




How important is the pre-training language (and domain)?

The Dutch model (final checkpoint) outperforms models with the same architecture trained on similar amounts of English data, or larger amounts of multilingual data (incl. Dutch), especially for dialogue speech (IFADV).

This is also reflected in downstream performance on automatic speech recognition: on average, the Dutch model has a 27% lower word error rate than the multilingual model.



Next steps

- Is the advantage of language-specific pre-training due to native-language effects (better encoding of Dutch-specific phones), or due to generally improved encoding of all Dutch speech sounds?
- What predictors best capture model phonetic and lexical learning trajectories? Do they follow human learning patterns?
- Is later-stage phonetic encoding (in the second peak) informed by higher levels of linguistic structure that peak in earlier layers?
- Does the encoding of higher-level linguistic structures causally affect model behaviour?