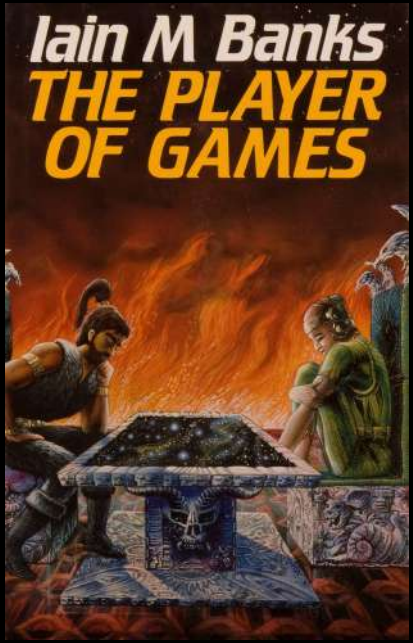
What can large spoken language models tell us about speech?

Is speech processing solved?

Herman Kamper

E&E Engineering, Stellenbosch University, South Africa

http://www.kamperh.com/

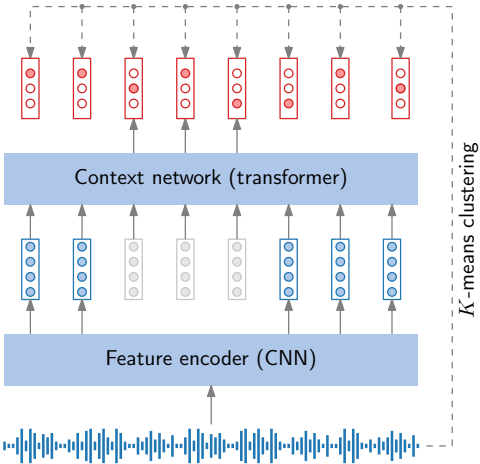

Iain M Banks
THE PLAYER OF GAMES

Supervised speech recognition and synthesis

i had to think of some example speech

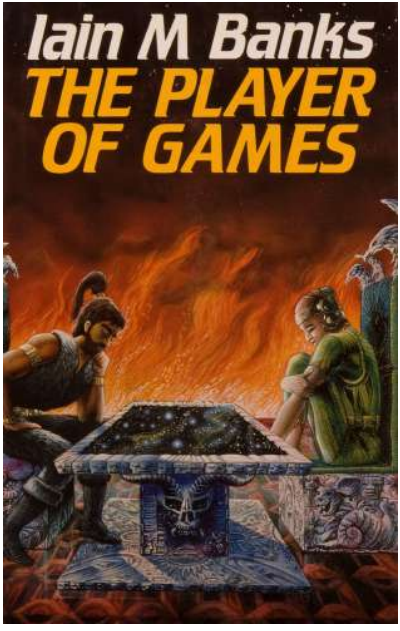
since speech recognition is really cool

Large spoken language models



- Models: CPC, wav2vec 2.0, HuBERT, WavLM
- Can now build automatic speech recognition systems with 10 min of data
- Low-resource text-to-speech
- Enabling textless language processing

Is speech processing solved?



1. Science: Understanding the speech signal better

Generative factors of speech

HH / Y / UW / M / ER

humour

Content: Discrete phonetic units

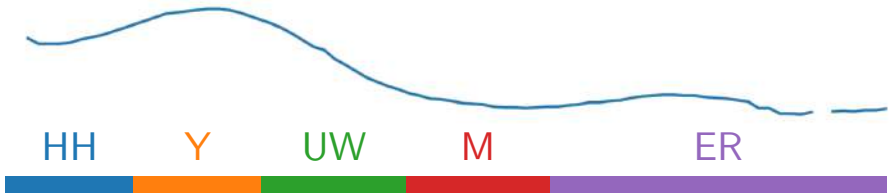
Slide and work credit: Benjamin van Niekerk

Generative factors of speech

Prosody: Rhythm

Slide and work credit: Benjamin van Niekerk

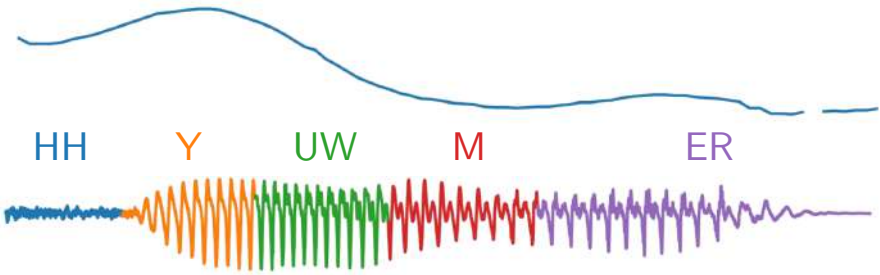
Generative factors of speech



Prosody: Rhythm, intonation, stresses

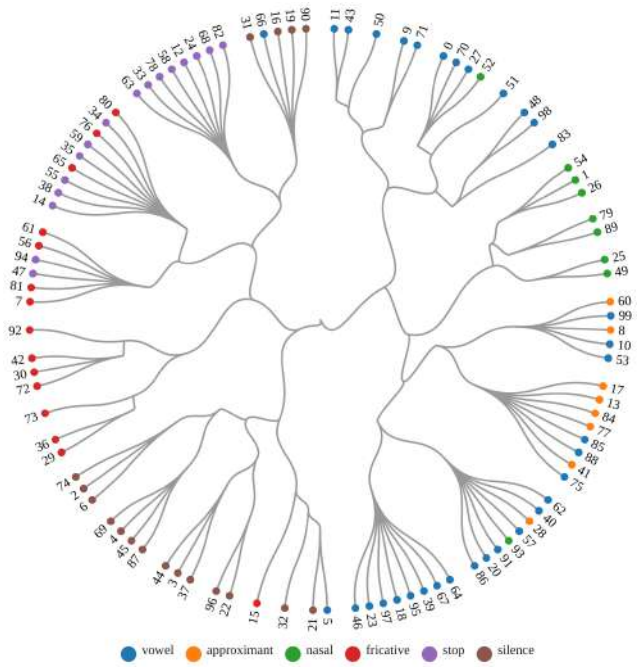
Slide and work credit: Benjamin van Niekerk

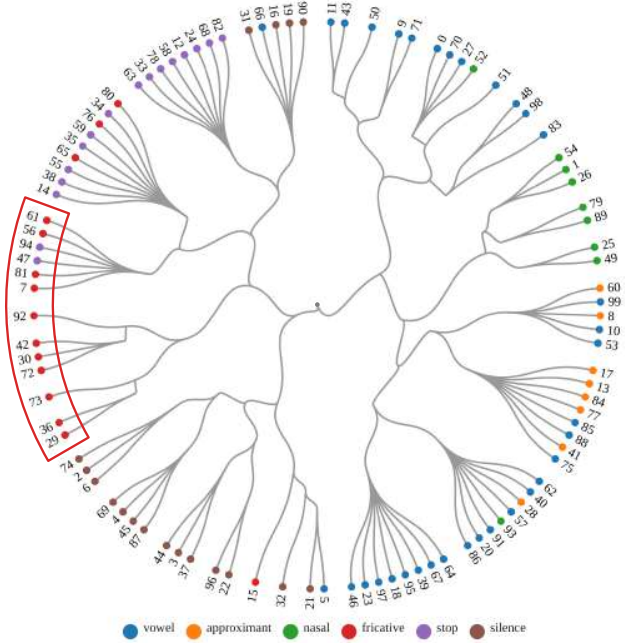
Generative factors of speech



Timbre (speaker characteristics), channel noise, etc.

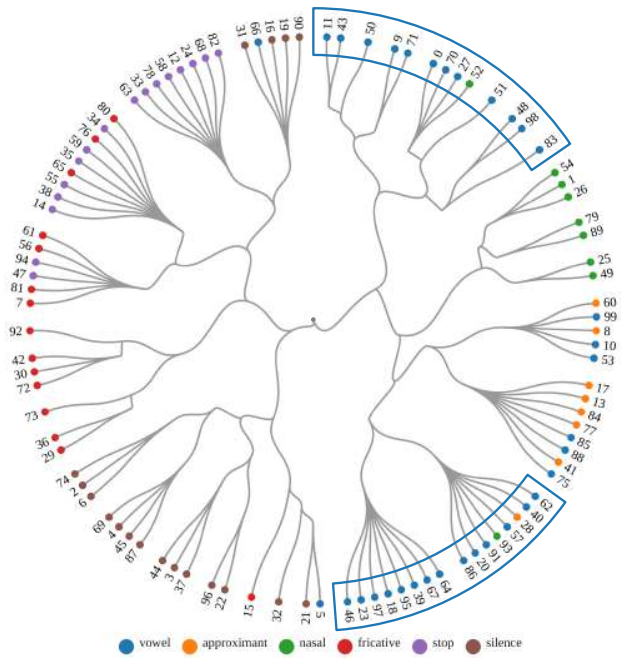
Slide and work credit: Benjamin van Niekerk


vowel
approximant
nasal
fricative
stop
silence



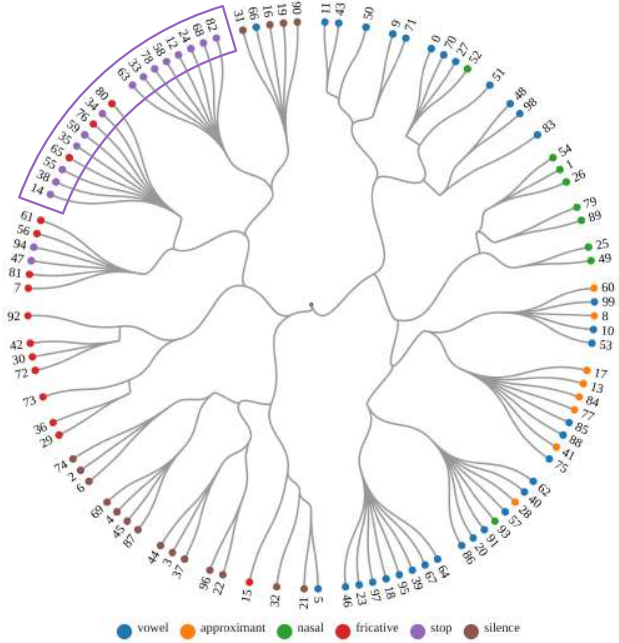
No modification: Play

Fricatives: Play



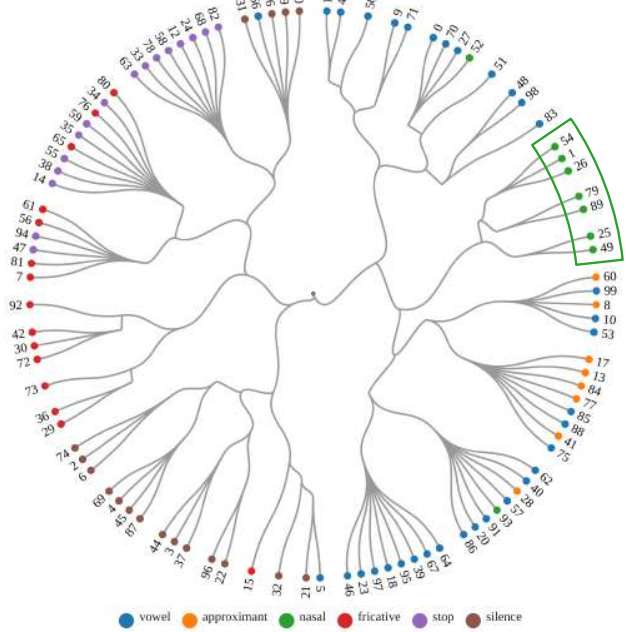
No modification: Play

Vowels: Play



No modification: Play

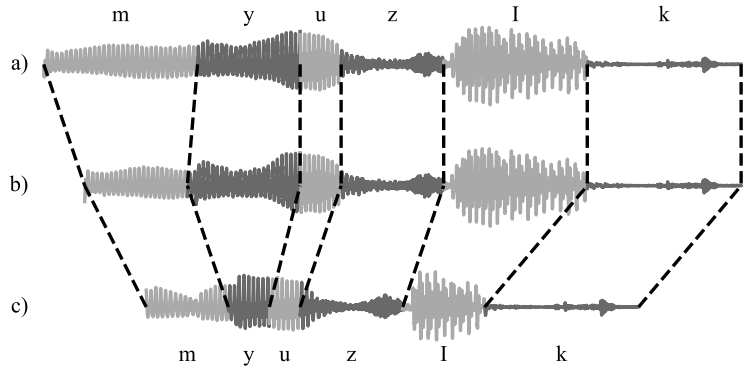
Stops: Play



No modification: Play

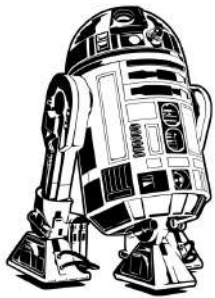
Nasals: Play

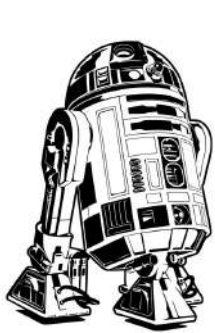
Unsupervised rhythm modelling for voice conversion

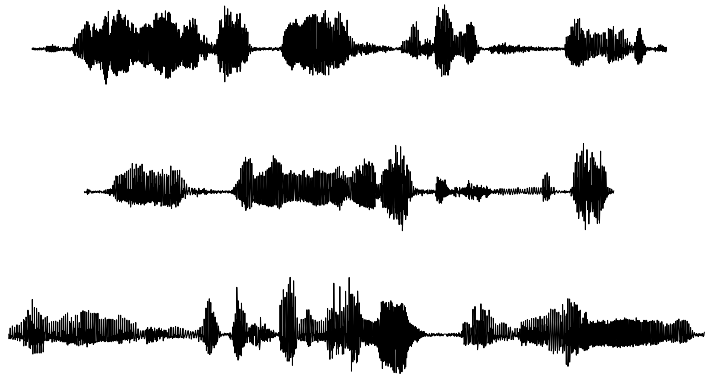


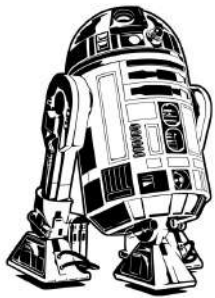
Input: Play Reference: Play Output: Play

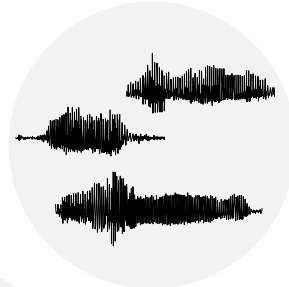
2. Science: Cognitive models of language acquisition

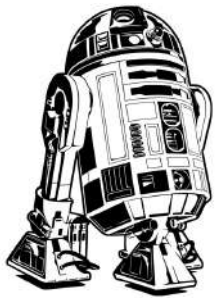


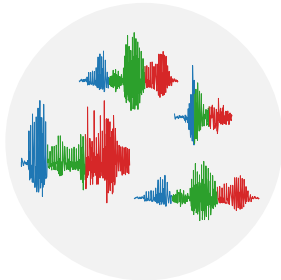


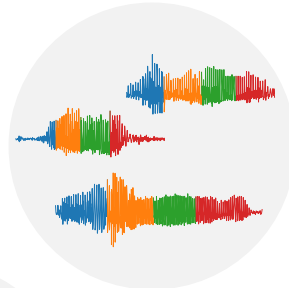




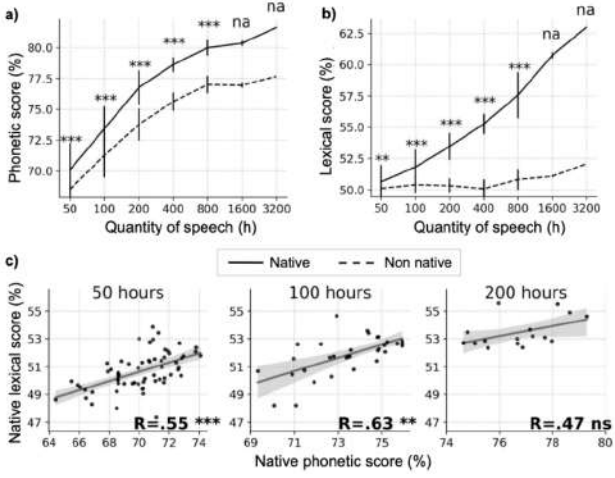






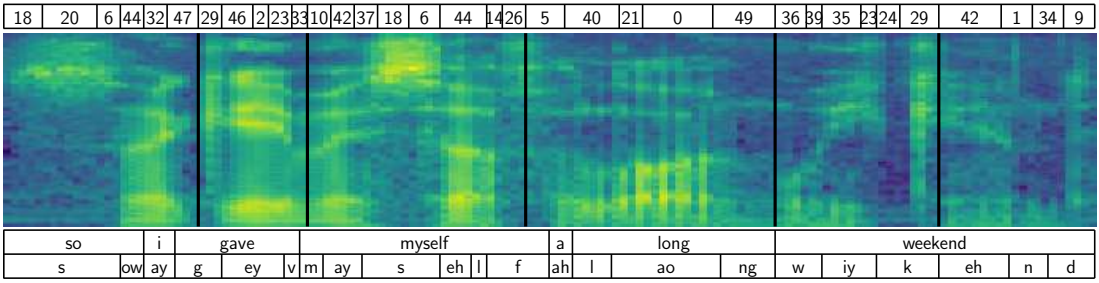


Contrastive predictive coding as a language learner



M. Lavechin et al., "Can statistical learning bootstrap early language acquisition? A modeling investigation," *PsyArXiv*, 2022.

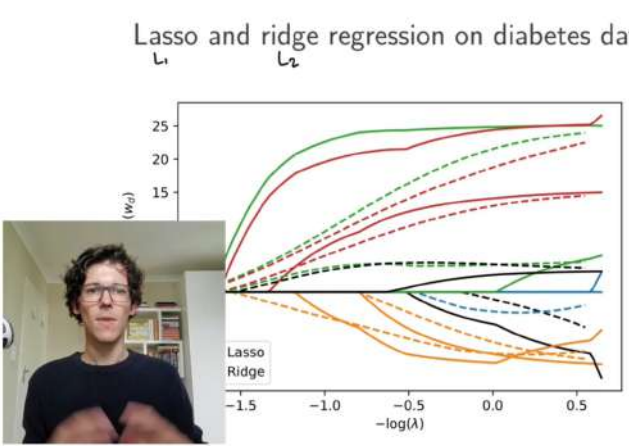
Unsupervised word segmentation



English cluster 1214: Play Xitsonga cluster 629: Play

3. Engineering: Tasks where we can't just do speech recognition

Expressive speech



Summary and conclusion

Large language models could help us:

1. Understand the speech signal better (science)
2. Model infant language acquisition (science)
3. Perform tasks that require more than speech recognition (engineering)
4. Solve useful tasks with speech recognition (engineering)

Want to join our group? Send me an email.

`http://www.kamperh.com/`