Do self-supervised speech models develop human-like perception biases?

Juliette MILLET and Ewan DUNBAR

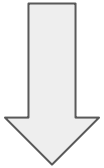
AAAI-SAS-2022, February 28th 2022

Self-supervised models

Great progress in speech recognition

Self-supervised models

Great progress in speech recognition

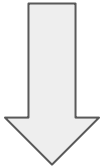


Thanks to self-supervised models
[1,2]

[1] Qiantong Xu et al. 2021. Self-training and pre-training are complementary for speech recognition. In ICASSP 2021
[2] Yu Zhang et al. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. arXiv preprint arXiv:2010.10504

Self-supervised models

Great progress in speech recognition

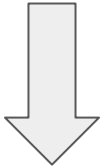


Thanks to self-supervised models

Self-supervised training

Self-supervised models

Great progress in speech recognition



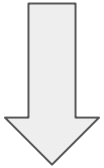
Thanks to self-supervised models

Self-supervised training

Fine-tuning on the task of speech recognition

Self-supervised models

Great progress in speech recognition



Thanks to self-supervised models

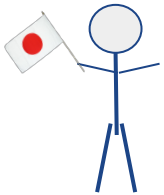
Self-supervised training

Fine-tuning on the task of speech recognition

Native language biases:

Native language biases:

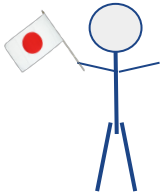
right/**l**ight



Japanese native speaker

Native language biases:

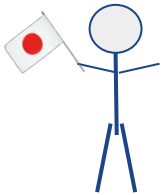
right/**l**ight



Japanese native speaker

tr**oo**p/tr**a**p

Native language biases:



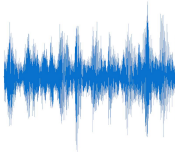
Japanese native speaker

English native speaker

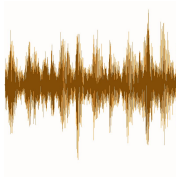
Human benchmark: Perceptimatic

https://docs.cognitive-ml.fr/perceptimatic/

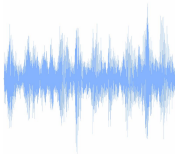
ABX test



A: ‘pop’



B: ‘pap’



X: ‘pop’

Results of a discrimination task:

- 4231 distinct test stimuli
- 662 phone contrasts
- 6 different languages
- 259 French-speaking participants
- 280 English-speaking participants

Do self-supervised speech models develop human-like perception biases?

Do self-supervised speech models develop human-like perception biases?

1) Do they reproduce human discrimination behaviour when exposed to the same language ?

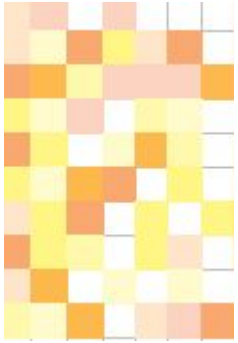
Do self-supervised speech models develop human-like perception biases?

1) Do they reproduce human discrimination behaviour when exposed to the same language ?
2) Does a change of training language has the same impact as a change of native language?

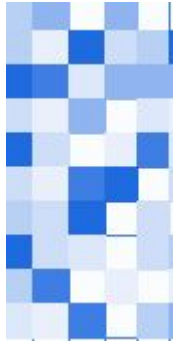
Models' evaluation

A: 'pop'

B: 'pap'



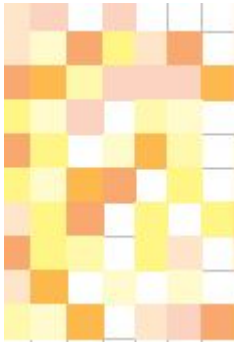
X: 'pop'



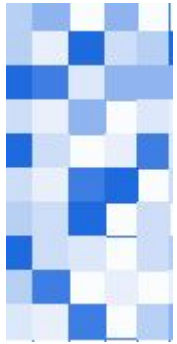
Models' evaluation

A: 'pop'

B: 'pap'



X: 'pop'



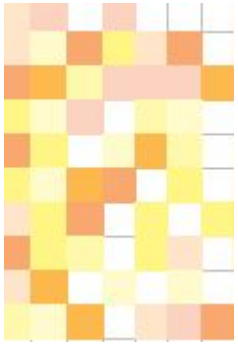
Dynamic time warping

Dynamic time warping

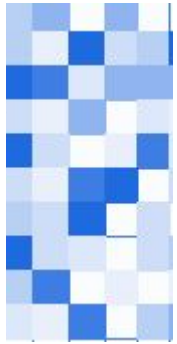
Models' evaluation

A: 'pop'

B: 'pap'



X: 'pop'



Dynamic time warping

Dynamic time warping

Discrimination result: Δ = DTW(wrong,X) - DTW(right,X)

Models to test

Self-supervised models:

[1] Morgane Rivière and Emmanuel Dupoux. 2021. Towards unsupervised learning of speech features in the wild. In 2021 IEEE Spoken Language Technology Workshop (SLT)

[2] Alexei Baevski et al.. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.arXiv preprint arXiv:2006.11477

[3] Wei-Ning Hsu et al. 2021.Hubert: How much can a bad teacher benefit asr pre-training? In ICASSP 2021

Models to test

Self-supervised models:

- CPC model (light): Contrastive Predictive Coding [1]

[1] Morgane Rivière and Emmanuel Dupoux. 2021. Towards unsupervised learning of speech features in the wild. In 2021 IEEE Spoken Language Technology Workshop (SLT)

[2] Alexei Baevski et al.. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.arXiv preprint arXiv:2006.11477

[3] Wei-Ning Hsu et al. 2021.Hubert: How much can a bad teacher benefit asr pre-training? In ICASSP 2021

Models to test

Self-supervised models:

- CPC model (light): Contrastive Predictive Coding [1]
- Wav2vec 2.0: Contrastive Predictive Coding using masking [2]

[1] Morgane Rivière and Emmanuel Dupoux. 2021. Towards unsupervised learning of speech features in the wild. In 2021 IEEE Spoken Language Technology Workshop (SLT)

[2] Alexei Baevski et al.. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.arXiv preprint arXiv:2006.11477

[3] Wei-Ning Hsu et al. 2021.Hubert: How much can a bad teacher benefit asr pre-training? In ICASSP 2021

Models to test

Self-supervised models:

- CPC model (light): Contrastive Predictive Coding [1]
- Wav2vec 2.0: Contrastive Predictive Coding using masking [2]
- HuBERT: teacher-student learning with a clustering goal [3]

[1] Morgane Rivière and Emmanuel Dupoux. 2021. Towards unsupervised learning of speech features in the wild. In 2021 IEEE Spoken Language Technology Workshop (SLT)

[2] Alexei Baevski et al.. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.arXiv preprint arXiv:2006.11477

[3] Wei-Ning Hsu et al. 2021.Hubert: How much can a bad teacher benefit asr pre-training? In ICASSP 2021

Models to test

Self-supervised models:

- CPC model (light): Contrastive Predictive Coding [1]
- Wav2vec 2.0: Contrastive Predictive Coding using masking [2]
- HuBERT: teacher-student learning with a clustering goal [3]

Trained on **English**, **French**

Models to test

Self-supervised models:

- CPC model (light): Contrastive Predictive Coding
- Wav2vec 2.0: Contrastive Predictive Coding using masking
- HuBERT: teacher-student learning with a clustering goal

Trained on **English**, **French** and **Acoustic scenes**

Models to test

Self-supervised models:

- CPC model (light): Contrastive Predictive Coding
- Wav2vec 2.0: Contrastive Predictive Coding using masking
- HuBERT: teacher-student learning with a clustering goal

Trained on **English**, **French** and **Acoustic scenes**

References:

- Baseline: MFCCs
- Topline: DeepSpeech [1]

[1] Dario Amodei et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning

Can self-supervised models predict human behaviour ?

Log likelihood of probit model

Similarity at the **stimuli** level

2 metrics

Can self-supervised models predict human behaviour ?

Log likelihood of probit model

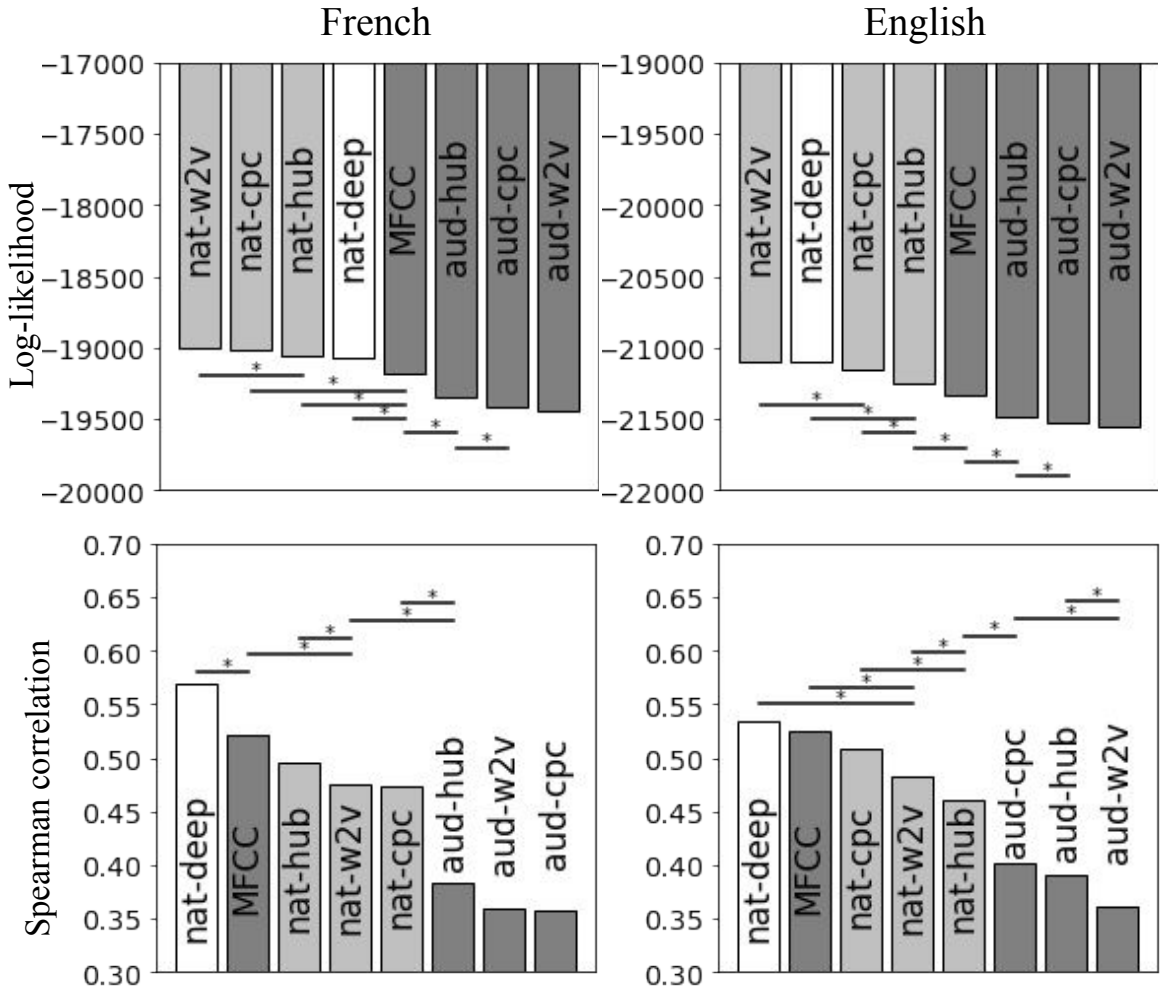
Similarity at the **stimuli** level

2 metrics

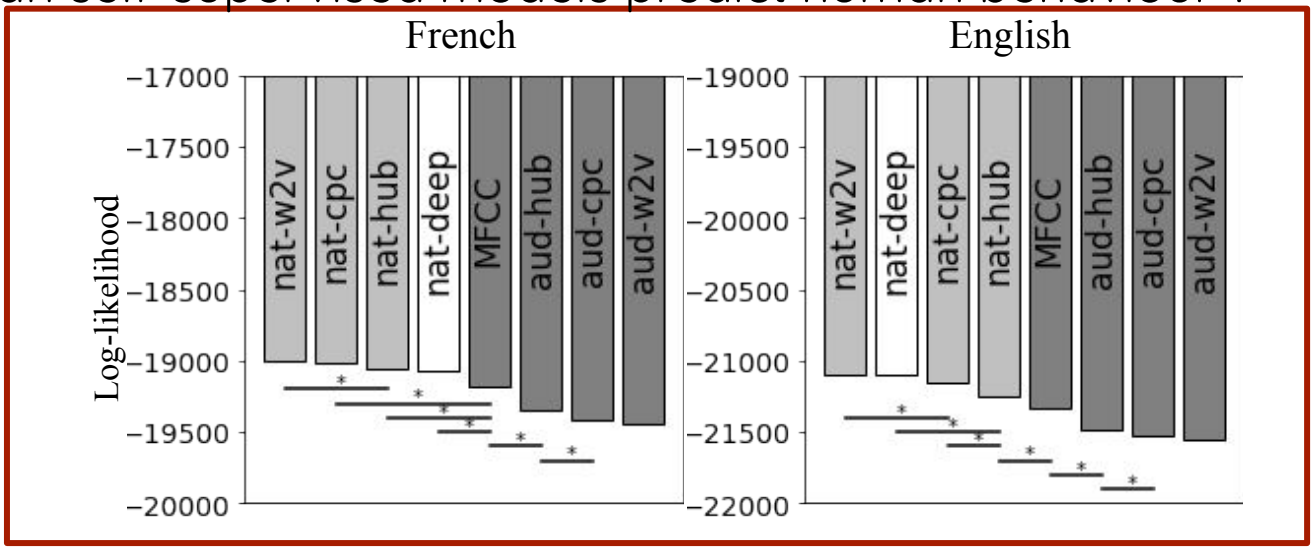
Spearman correlation

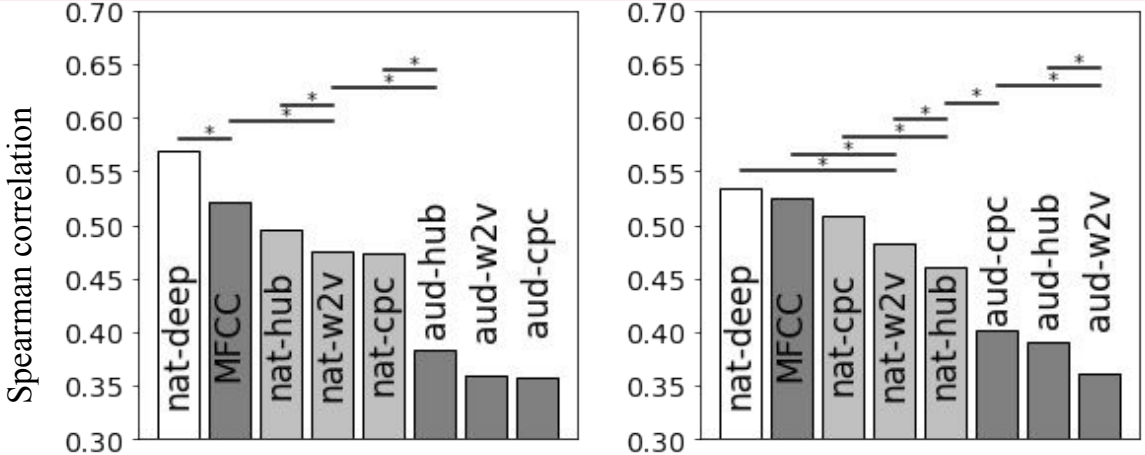
Similarity at the **contrast** level

Results: Can self-supervised models predict human behaviour ?



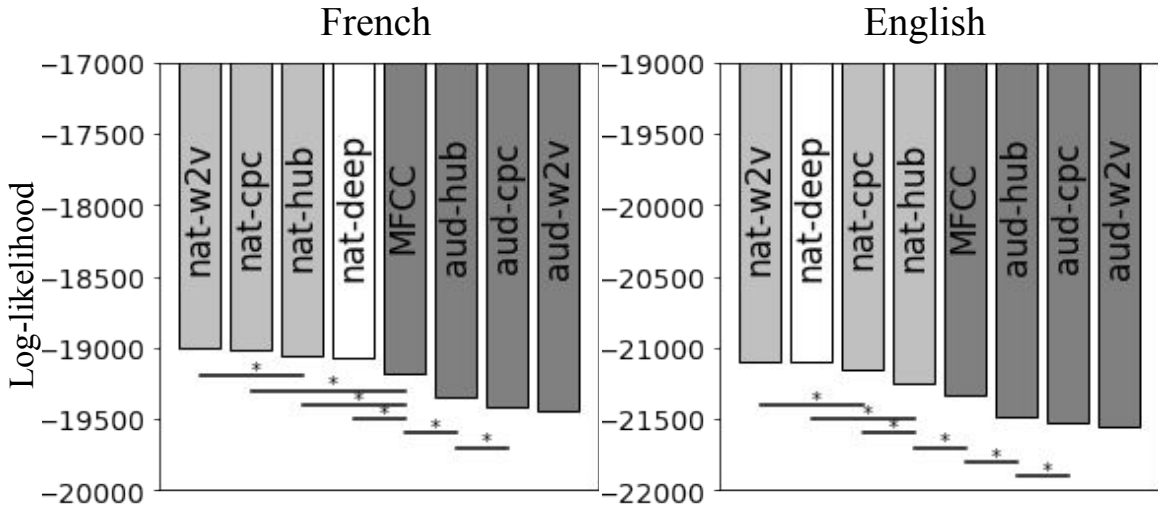
Results: Can self-supervised models predict human behaviour ?

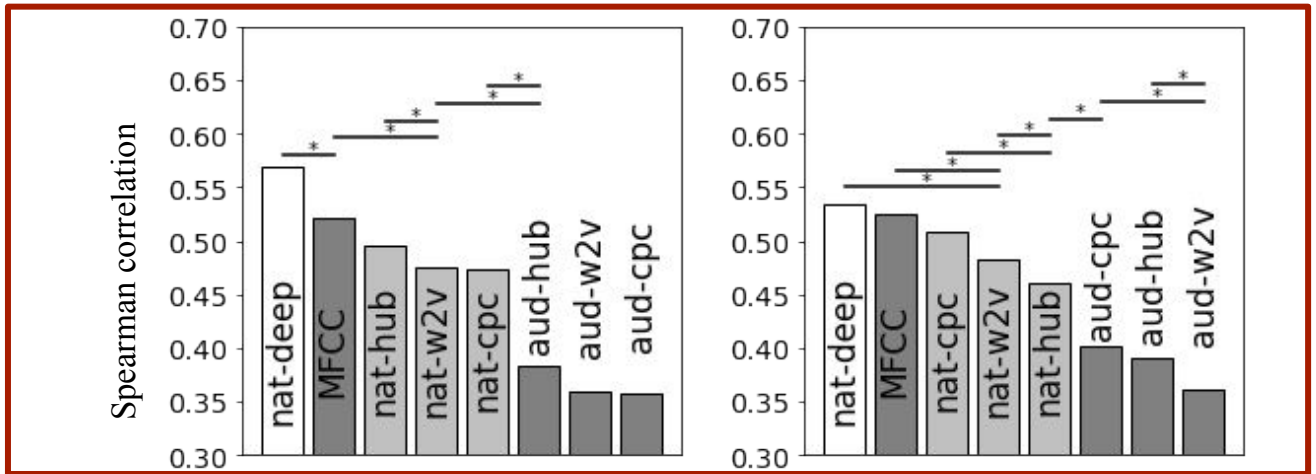




Stimuli level

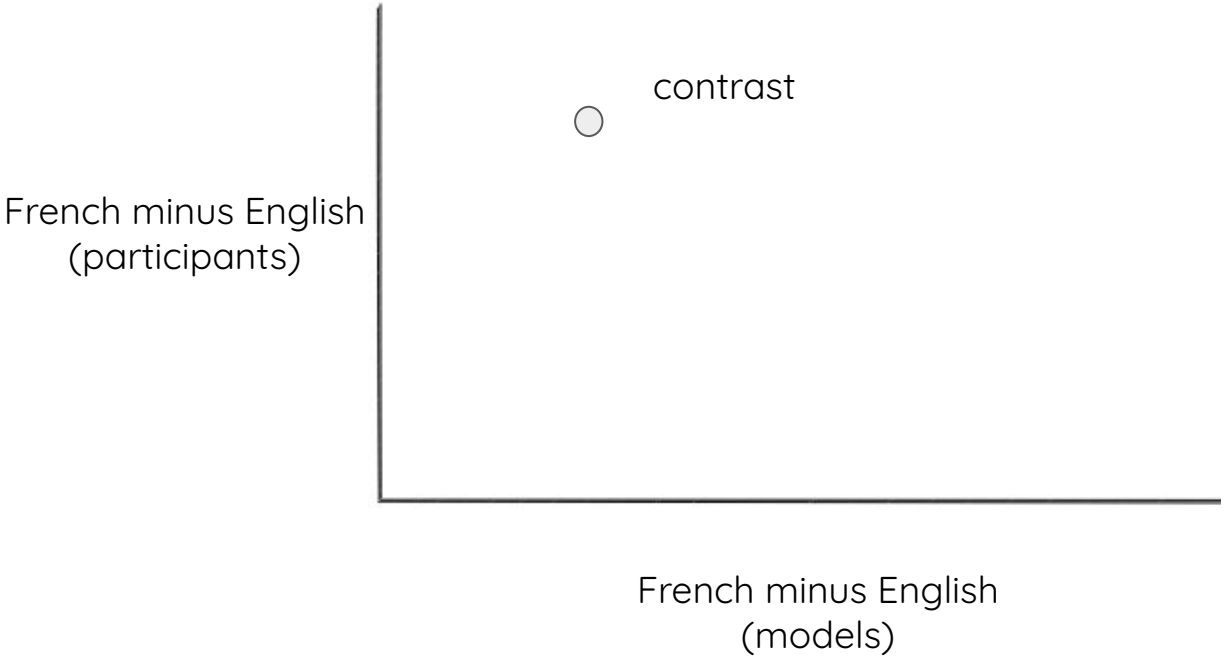
Results: Can self-supervised models predict human behaviour ?



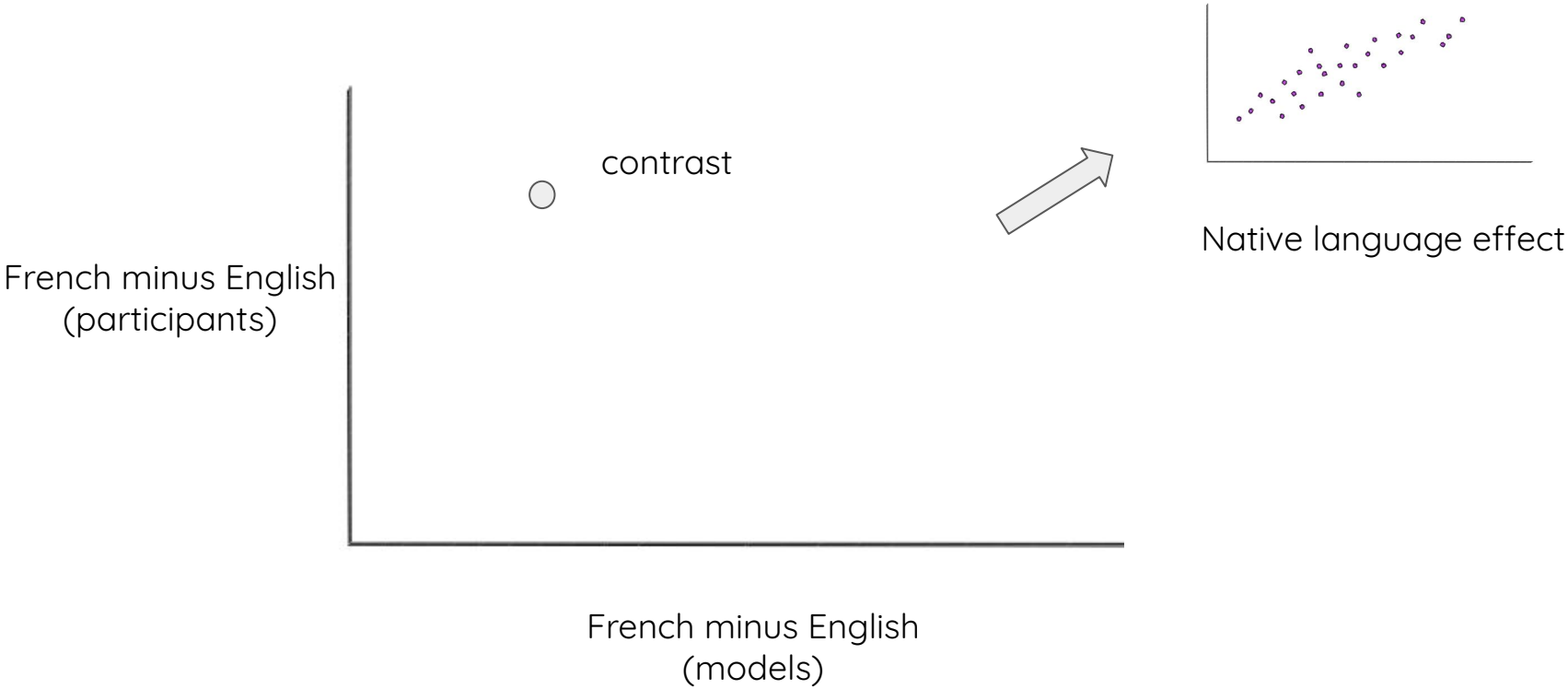


Contrast level

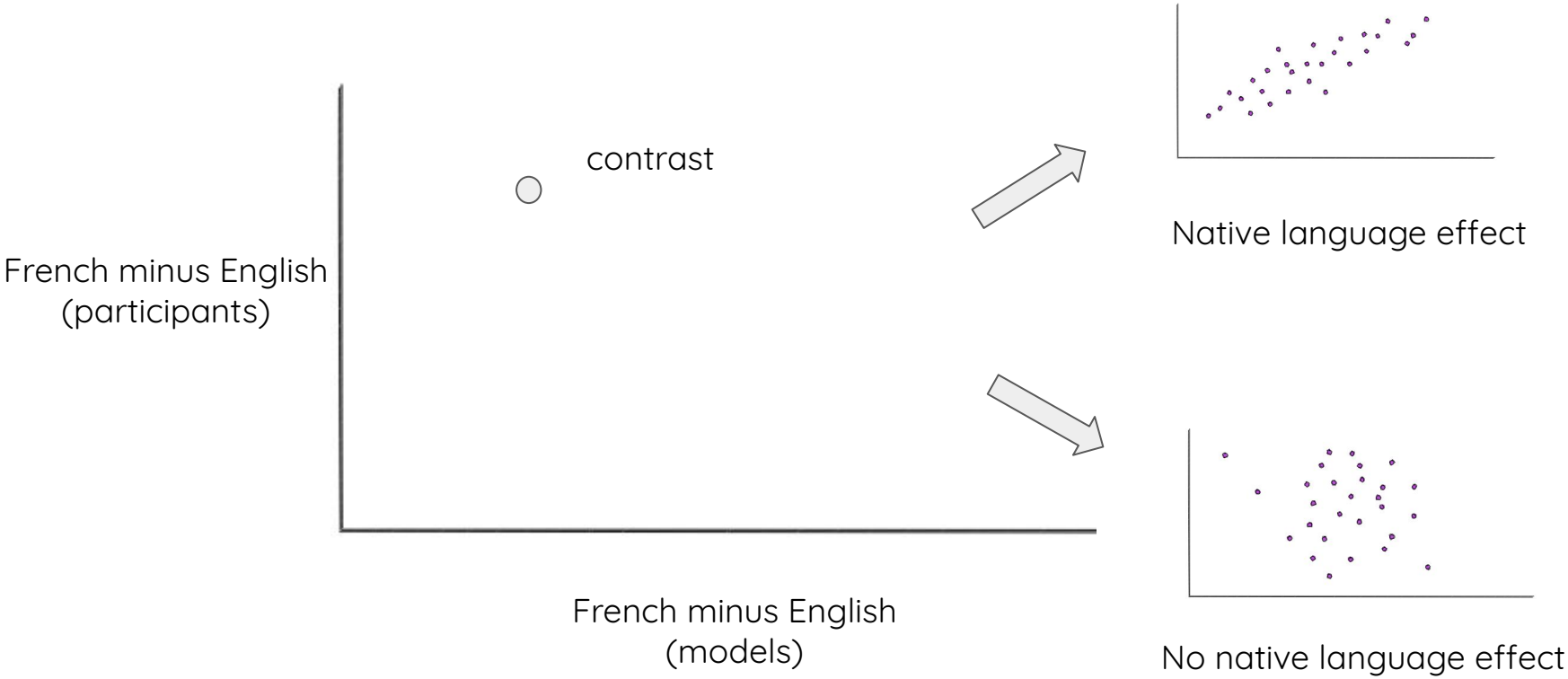
Do they reproduce **differences** in human behaviour?



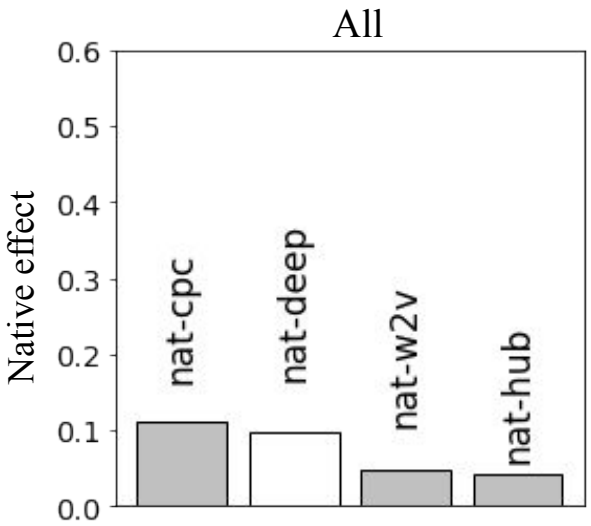
Do they reproduce **differences** in human behaviour?



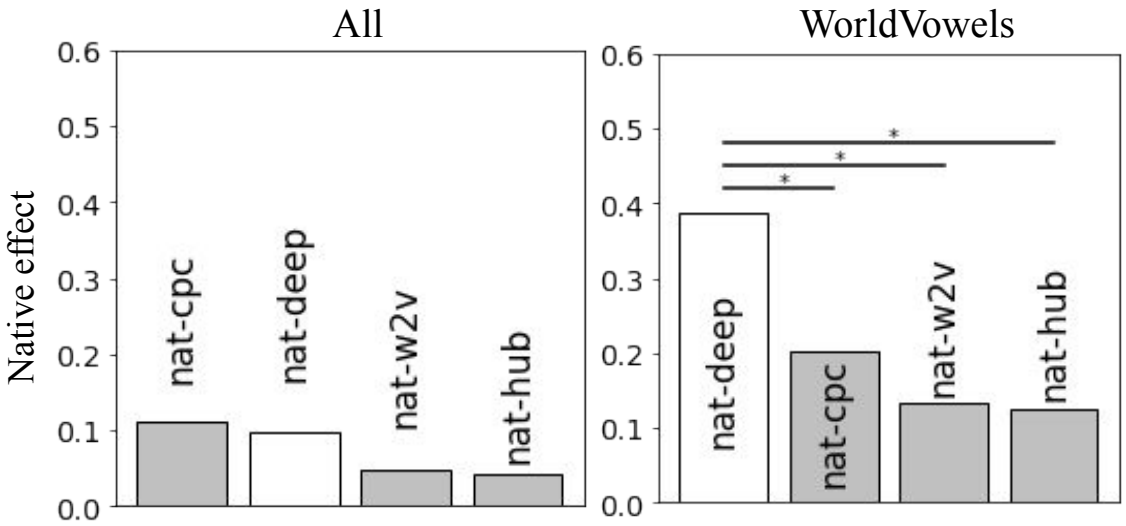
Do they reproduce **differences** in human behaviour?



Results: Do they reproduce differences in human behaviour?



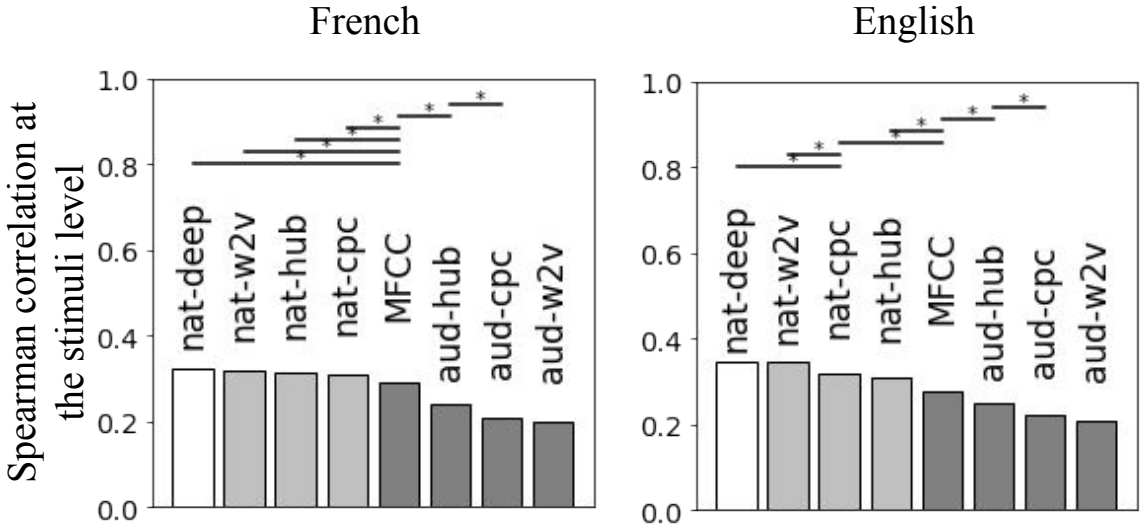
Results: Do they reproduce differences in human behaviour?

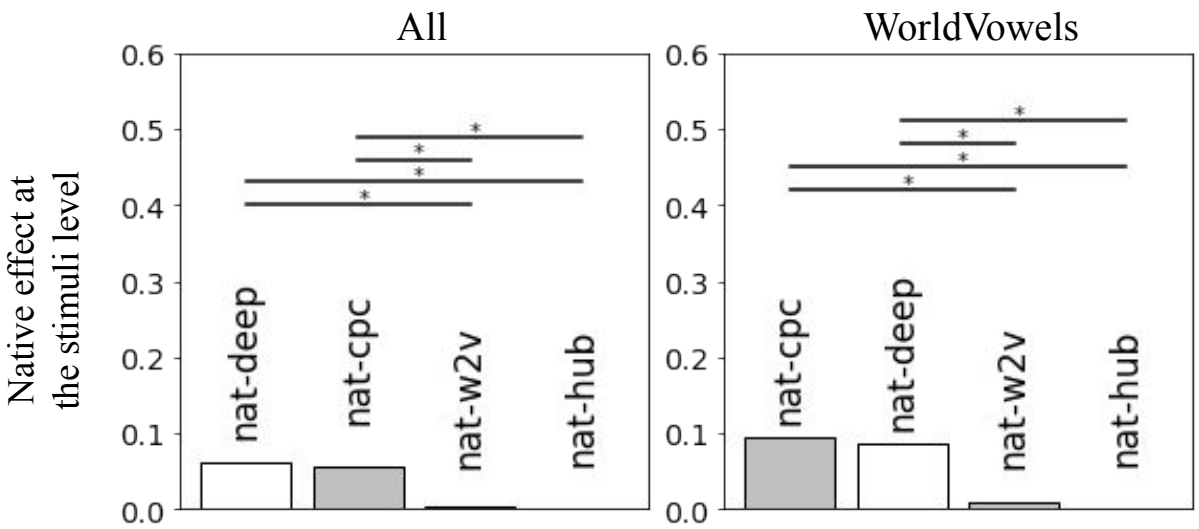


Conclusion

- Self-supervised models can predict human discrimination behaviour at the **stimuli level** but not very well at the **contrast level**
- They need to be trained on **speech**
- They show a very small native language effect, except for CPC

Thanks ! Questions ?


French
English
Spearman correlation at the stimuli level
1.0
0.8
0.6
0.4
0.2
0.0
nat-deep
nat-w2v
nat-hub
nat-cpc
MFCC
aud-hub
aud-cpc
aud-w2v
*


All
WorldVowels
Native effect at
the stimuli level
0.6
0.5
0.4
0.3
0.2
0.1
0.0
nat-deep
nat-cpc
nat-w2v
nat-hub
*
*
*
*
nat-cpc
nat-deep
nat-w2v
nat-hub
*
*
*
*