APPENDIX

Detailed losses used by the models

The loss used by the **CPC** model is the following:

$$\mathcal{L}_t = -\frac{1}{K}\sum_{k=1}^{K}\log\left[\frac{\exp\left(\phi\left(\mathbf{x}_{t+k}\right)^{\top}\mathbf{A}_k\mathbf{z}_t\right)}{\sum_{\mathbf{n}\in\mathcal{N}_t}\exp\left(\phi(\mathbf{n})^{\top}\mathbf{A}_k\mathbf{z}_t\right)}\right] \quad (2)$$

Where A_k is a learned linear classifier, ϕ is the encoder, and N_i is the set of negative examples. With an input x_t and an output $\mathbf{z}_t = \psi\left(\phi\left(\mathbf{x}_1\right),\ldots,\phi\left(\mathbf{x}_t\right)\right)$, with ψ the sequential model, it pushes the model to identify the K next outputs $\phi\left(\mathbf{x}_{t+k}\right)$ in the future, in comparison with randomly sampled outputs from another part of x.

The loss used by the **wav2vec 2.0** model is the following:

$$\mathcal{L}_m = -\log\frac{\exp\left(\mathrm{sim}\left(\mathbf{c}_t,\mathbf{q}_t\right)/\kappa\right)}{\sum_{\tilde{\mathbf{q}}\sim\mathbf{Q}_t}\exp\left(\mathrm{sim}\left(\mathbf{c}_t,\tilde{\mathbf{q}}\right)/\kappa\right)} \quad (3)$$

for a masked time step t, the model has to choose the true quantized speech representation q_t in a set of $K+1$ quantized candidate representations $\tilde{\mathbf{q}} \in \mathbf{Q}_t$ which includes $\mathbf{q}_t$ and K distractors. The model also use a diversity loss so the representation in the quantizer dictionary be as diverse as possible, for more details, see (Baevski et al. 2020).

The loss used by **HuBERT** is the following:

$$L(f;X,M,Z) = \alpha\sum_{t\in M}\log p_f\left(z_t \mid \tilde{X},t\right) + (1-\alpha)\sum_{t\notin M}\log p_f\left(z_t \mid X,t\right)$$

With $\alpha \in [0,1]$, M the set of masked frames, f the cluster assignment predictor, and $\hat{X}$ masked frames.

Predicting human results: results on sub-datasets

We present the results on the different Perceptimatic subsets. The results for Cogsci 2019 can be seen in Figure 5, for WorldVowels in Figure 4, for Zerospeech in Figure 8, for pilot-july in Figure 6, and for pilot-august in Figure 7. These results should be taken carefully, in particular for the Cogsci subset and the pilots, as not much contrasts and stimuli were tested for these subsets compared to the others.

Difference in ABX score between French and English models

To complete Table 1, we present in Figure 9 the detailed ABX score difference between a native discrimination setting (English models and participants discriminating English contrasts and same for French) and a non-native discrimination setting (English models and participants discriminating French contrasts and vice-versa). Humans' ABX scores differences show that English-speaking participants are not always better than French-speaking participants at discriminating English sounds (for the Zerospeech subsets for example).

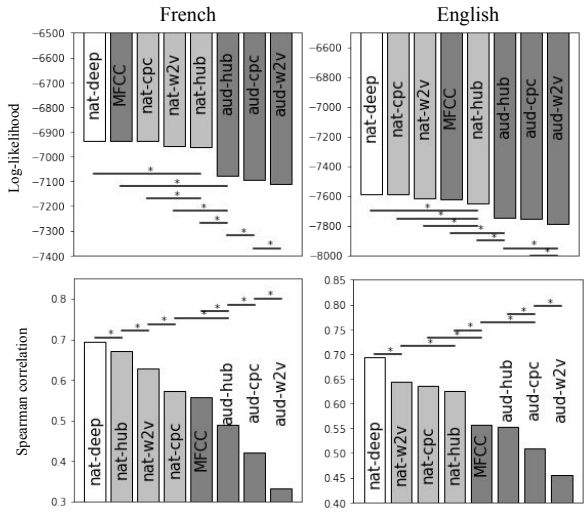

Figure 4: Results on the **WorldVowels** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the self-supervised model trained on speech recordings (in light grey), and the baselines in darker grey (neutral acoustic features and models trained on acoustic scenes).

Language preference

We assess whether models in the native training condition predict discriminability better than the corresponding models in the non-native training condition. Figure 10 plots the subtraction of the $\ell\ell$ and ρ scores in the non-native setting from the corresponding scores in the native setting (across the entire Perceptimatic dataset).

For both the (experimental item-level) $\ell\ell$ and the (phone contrast-level) ρ score, DeepSpeech consistently outperforms over wav2vec 2.0. This is in contrast with the overall prediction performance reported above, where wav2vec 2.0 was on par with DeepSpeech, DeepSpeech generally shows a relative advantage for predicting the behaviour of listeners whose native language is the same as the training language, while wav2vec 2.0 does not.

There is a striking difference between languages in the performance of DeepSpeech: for English, the native DeepSpeech shows a substantial advantage over the non-native (French-trained) DeepSpeech which is not present for the French datasets. Similarly, in French, the native HuBERT shows an advantage over the non-native (English-trained) HuBERT, while the reverse is true in English. However, these two major differences may be in part explained by global effects: the French-trained HuBERT model is better at predicting the results for all participants (not just French-speaking participants), as is the English-trained DeepSpeech model.

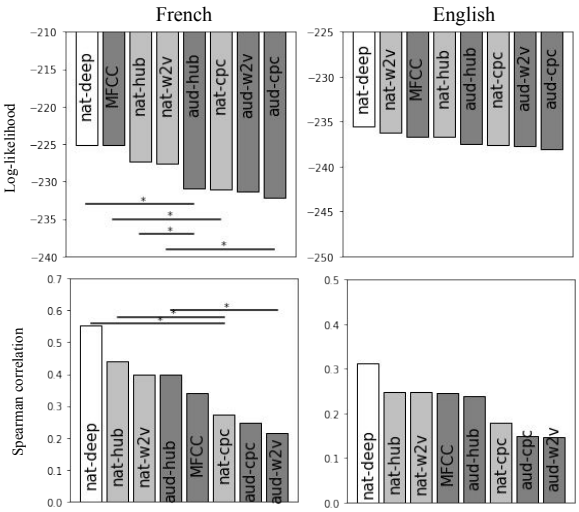

Figure 5: Results on the **Cogsci-2019** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the self-supervised model trained on speech recordings (in light grey), and the baselines in darker grey (neutral acoustic features and models trained on acoustic scenes).

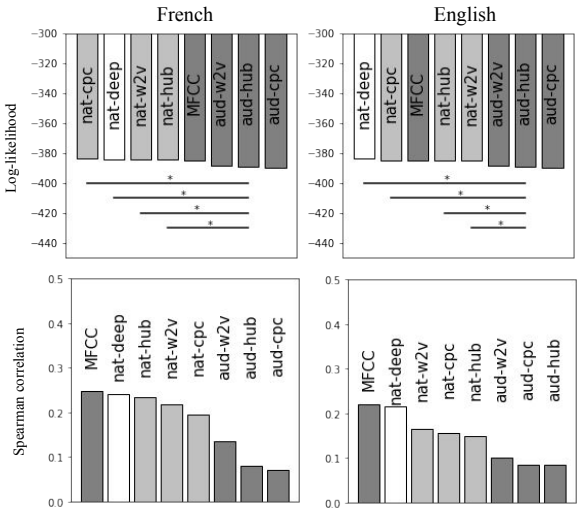

Figure 6: Results on the **pilot-july-2018** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the self-supervised model trained on speech recordings (in light grey), and the baselines in darker grey (neutral acoustic features and models trained on acoustic scenes).

Using pretrained models on more data

We compare our models with pretrained models available online. For English, we tested a wav2vec and a HuBERT model trained on Librispeech (Panayotov et al. 2015) (960 h) and for French, we tested a wav2vec model trained on the French Voxpopuli dataset (Wang et al. 2021) (4.5k h). The results of these models compared to ours and MFCCs can be seen in Figure 11. Their different ABX scores can also be seen in Table 2. Models trained on English are evaluated on English-speaking participants (and English contrast for the ABX scores), and same for French.

Models	Zerospeech		WorldVowels		PA
	FR	EN	FR	EN	EN
w2v-nat	**0.88**	0.88	0.71	0.83	0.84
w2v-pret	0.85	0.86	0.69	0.84	0.86
hub-nat	0.87	0.87	**0.76**	0.83	0.82
hub-pret	-	**0.89**	-	**0.89**	**0.90**
mfccs	0.76	0.77	0.73	0.76	0.88

Table 2: ABX scores of our self-supervised models (-nat) compared to pretrained ones (-pret). Best results for each subset is in bold

Testing DeepSpeech using orthographic transcriptions

We tested two kinds of supervised references: one trained to produce phonemic transcriptions (the one used in the main article) and another trained to produce orthographic transcriptions. In general, training on phonemic transcriptions led the internal representations of the model to be closer to humans' perceptual space, as it can be seen in Figure 12. A comparison of English-speaking participants' discrimination ability and the two supervised models' Δ-values can also be seen in Figure 13. Models trained on phonemic transcriptions are better at predicting human behaviour than the ones trained on orthographic transcriptions. These results highlight on the one hand the impact of the labels used during supervised training, which can lead to non human-like speech representational space, and on the other hand the fact that humans probably use informations more similar to phoneme categories than possible orthographic transcriptions during a discrimination task.

The amount of training data may also play a role, as large training sets could lead to "overfitting," in a loose sense, to fine "superhuman" acoustic details of phone classification. Previous sections shows that training size does *not* have this effect on the self-supervised models studied here. We leave analysis of the supervised case for future work.

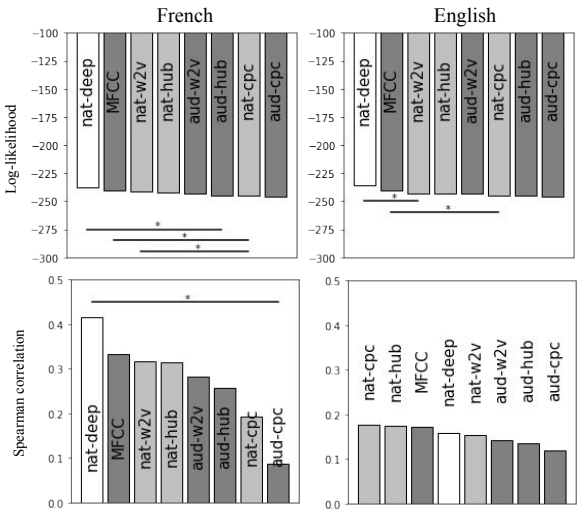

Figure 7: Results on the **pilot-august-2018** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the self-supervised model trained on speech recordings (in light grey), and the baselines in darker grey (neutral acoustic features and models trained on acoustic scenes).

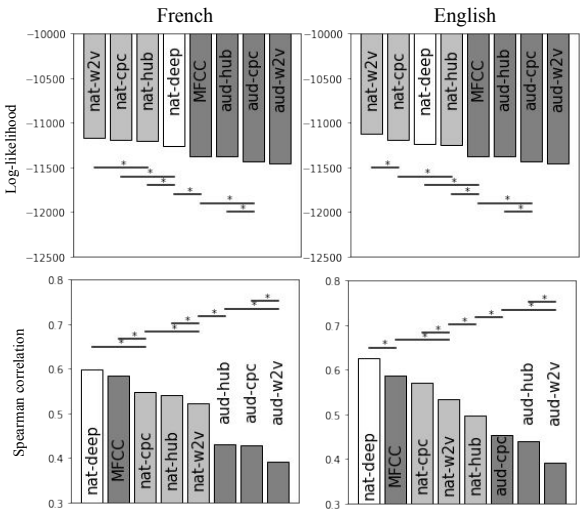

Figure 8: Results on the **Zerospeech** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the self-supervised model trained on speech recordings (in light grey), and the baselines in darker grey (neutral acoustic features and models trained on acoustic scenes).

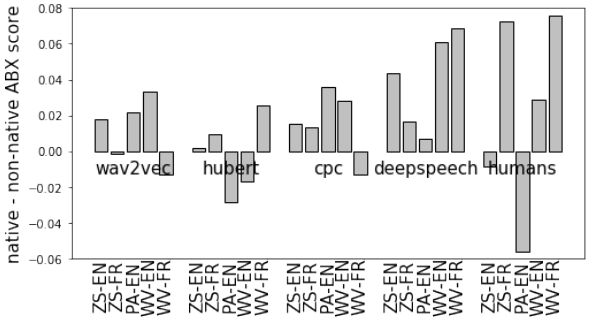

Figure 9: ABX score difference between native setting and non-native setting for the different models tested. The bigger the bar above zero, the bigger difference.

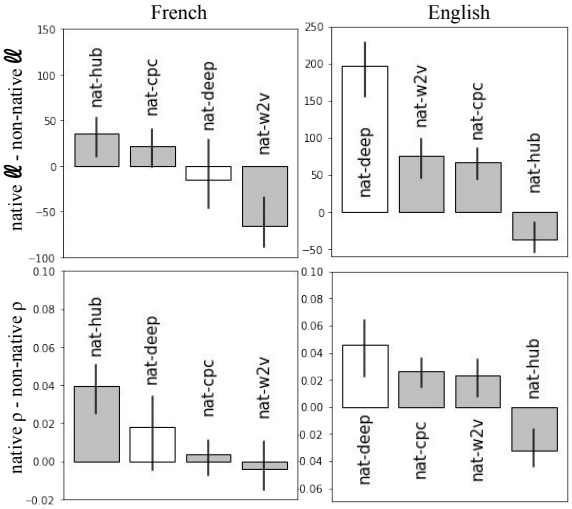

Figure 10: Native minus non-native log-likelihood values (top) and Spearman correlations (bottom) for French (*left*) and English participants (*right*). The higher the bar above zero, the better the native setting is compared to the non-native setting. The supervised reference is in white, the self-supervised models are in light grey. Black lines indicate 95% confidence intervals.

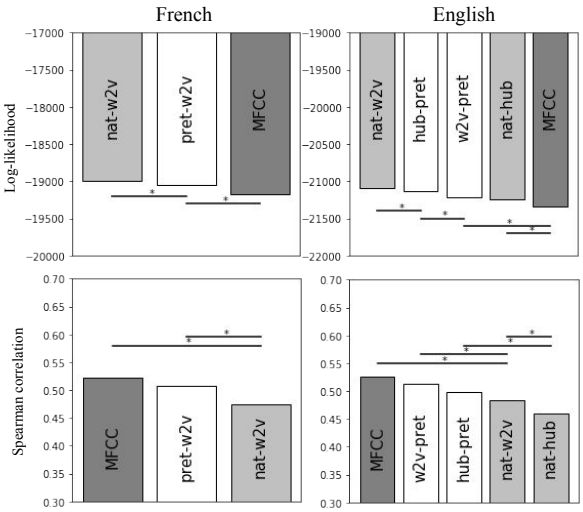

Figure 11: Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The pretrained models are in white to distinguish it from our self-supervised models trained on only 600h of speech.

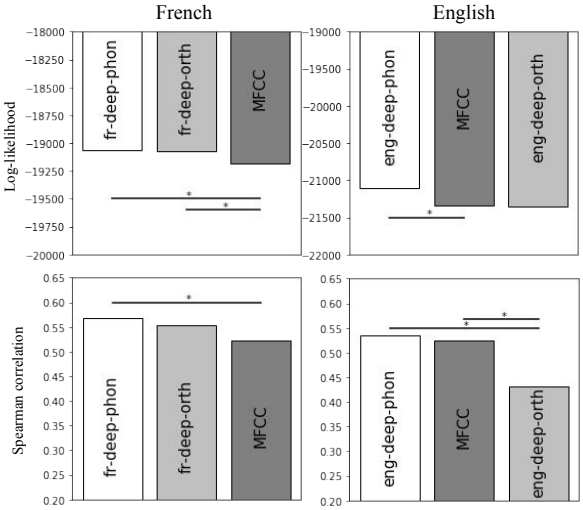

Figure 12: Results of Deepspeech trained on phonemic transcriptions (phon) or orthographic (orth), compared with MFCCs. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant.

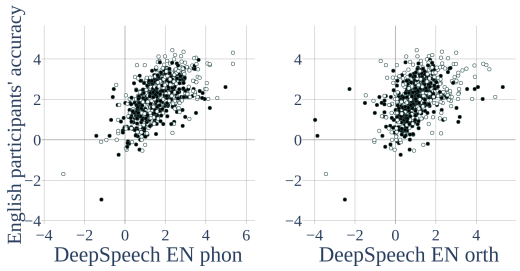

Figure 13: Average of English listeners' results (higher: better discrimination) against average δ from (**left**) supervised reference trained on phonemic transcriptions (**right**) trained on orthographic transcriptions. Each point is a contrast. Measures are normalized by dividing by standard deviation over the entire data set. Black circles are non-native contrasts, white ones are native (English).