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

Juliette Millet
CoML, ENS/CNRS/EHESS/INRIA/PSL,
LLF, University of Paris, CNRS,
CRI, FAN, IIFR, University of Paris,
Paris, France
juliette.millet@cri-paris.org

Ewan Dunbar
CoML, ENS/CNRS/EHESS/INRIA/PSL,
Paris, France
University of Toronto, Toronto, Canada
ewan.dunbar@utoronto.ca

Abstract

Self-supervised models for speech processing form representational spaces without using any external labels. Increasingly, they appear to be a feasible way of at least partially eliminating costly manual annotations, a problem of particular concern for low-resource languages. But what kind of representational spaces do these models construct? Human perception specializes to the sounds of listeners' native languages. Does the same thing happen in self-supervised models? We examine the representational spaces of three kinds of state-of-the-art self-supervised models: wav2vec 2.0, HuBERT and contrastive predictive coding (CPC), and compare them with the perceptual spaces of French-speaking and English-speaking human listeners, both globally and taking account of the behavioural differences between the two language groups. We show that the CPC model shows a small native language effect, but that wav2vec 2.0 and HuBERT seem to develop a universal speech perception space which is not language specific. A comparison against the predictions of supervised phone recognisers suggests that all three self-supervised models capture relatively fine-grained perceptual phenomena, while supervised models are better at capturing coarser, phone-level, effects of listeners' native language, on perception.

1 Introduction

Recent advances in speech recognition and representation learning show that self-supervised pre-training is an excellent way of improving performance while reducing the amount of labelled data needed for training. For example, for the LibriSpeech dataset (Panayotov et al., 2015), the current best word error rates (Xu et al., 2021; Zhang et al., 2020) are obtained by systems based on the self-supervised wav2vec 2.0 model (Baevski et al., 2020). Systems using self-supervised pre-training, both using wav2vec 2.0 and using HuBERT (Hsu et al., 2021a,b), show excellent word error rates after having been fine-tuned on only ten minutes of labelled data.

What is the effect of this self-supervised pre-training? What type of representational spaces are learned by these models? Lakhotia et al. (2021) compared wav2vec 2.0, HuBERT, and contrastive predictive coding (CPC: Oord et al. 2017; Rivière and Dupoux 2021) using an ABX discriminability metric (Schatz, 2016), demonstrating that all three models preserve and enhance linguistically relevant speech sound contrasts in the language they are trained on. We build on this work, asking how these representational spaces compare to the perceptual spaces of human listeners, as inferred from behaviour on phone discrimination experiments.

Human listeners develop speech perception biases under the influence of their native languages. For example, Japanese native speakers tend to confuse the English sounds /r/ and /l/ (Yamada and Tohkura, 1990) (***r**ight* and ***l**ight* in English will be perceived as the same or very similar), and English native speakers struggle with the French contrast /y/-/u/ (Levy, 2009), having difficulty perceiving the difference between words such as ***ru**e* (/y/: "street") and ***rou**e* (/u/: "wheel"). These misperceptions start to show early on in the native language acquisition process: infants older than 6 months exhibit a facilitating effect at discriminating sounds from their native language, but a decline at doing so for some non-native sounds (Kuhl et al., 2006). As the importance of this improvement for native sounds and this decline for non-native sounds seems to have a positive impact on infants' future language ability (Tsao et al., 2004; Kuhl et al., 2005), having a perceptual space with native language biases is probably essential to perceive and understand correctly native speech in all situations (with environmental noises, speaker change, etc). If our goal is to have speech models that are as resilient and as adaptable as humans, it is thus

interesting to see if they present the same native language specific biases.

By measuring human listeners' ability to discriminate a variety of familiar and unfamiliar speech sounds, we can create a detailed profile of listeners' perceptual biases in the form of a set of sounds' discriminabilities. We then ask whether the training language influences self-supervised speech models in the same way that human listeners' native languages do.

In order to study speech models' perception biases and compare them with humans', we use the Perceptimatic benchmark datasets,[1] a collection of experimental speech perception data intended to facilitate comparison with machine representations of speech. As of this writing, Perceptimatic contains French- and English-speaking participants' behaviour on discrimination tasks for phones in six different languages, for a total of 662 phone contrasts, along with the sound stimuli used during the experiments.

As in Lakhotia et al. (2021), we test state-of-the-art self-supervised models: wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021a,b) and a CPC model (Rivière and Dupoux, 2021). We train these models on English and French speech recordings (the native languages of the participants in Perceptimatic). We compare the performance of these self-supervised models with a supervised ASR model, DeepSpeech (Amodei et al., 2016), trained on the same data but using phonemic labels. To study the degree to which the models' representational space is impacted by properties of speech per se, we also train the same models on recordings of acoustic scenes not including human vocalisations (environmental noises, animal sounds, music, and so on). We use mel-frequency cepstrum coefficients (MFCCs) as an acoustic baseline.

We show that: (1) Self-supervised models trained on speech recordings are better than models trained on acoustic scenes (non-speech) to discriminate speech sounds and to predict human discrimination behaviour (2) They are good at predicting human discrimination behaviour at the stimuli level, but they are worse than neutral acoustic features when we average human results per contrast (3) They show very few native (training) language effect.

All our code and data are freely available.[2]

[1] https://docs.cognitive-ml.fr/perceptimatic/

[2] https://github.com/JAMJU/Sel_supervised_models_perception_biases

2 Related work

We are not the first to compare speech models' representational spaces with humans. Feather et al. (2019) used metamers as a tool to compare deep neural networks with humans. In a comparison between three speech recognition models, including a fine-tuned wav2vec 2.0 model, Weerts et al. (2021) showed that wav2vec 2.0 was the best at matching human low-level psycho-acoustic behaviour. However, the model exhibited clear differences with respect to humans—showing, for example, heightened sensitivity to band-pass filtering and an under-reliance on temporal fine structure.

To perform a comparison at a slightly higher level of speech perception, Scharenborg et al. (2018) visualised a supervised ASR model's internal representations of different speech sounds to investigate its adaptation to new ambiguous phone categories and compare it to humans' behaviour.

Multiple datasets containing human behavioural data have been collected and openly released to encourage comparison of models with humans. It is for this reason that the Interspeech 2008 Consonant Challenge (Cooke and Scharenborg, 2008) and the OLLO database (Meyer et al., 2010), containing humans' phone identification behaviour in different paradigms, were created. This is also the case for the datasets making up the Perceptimatic database (Millet et al., 2019; Millet and Dunbar, 2020a,b; Millet et al., 2021) that we employ in this article, which were individually used to study less well-performing models than the ones we use here.

More than just informing us on the kind of information speech models learn, comparing them with humans can have a broader impact on our knowledge of how human perceive speech, and how they learn to do so. Schatz et al. (2021) showed, for example, that a simple self-supervised speech model reproduces the reduced sensitivity to the English [r]/[l] contrast when trained on Japanese speech recordings. Pointing to the fact that the model used lacks abstract phone categories, the authors proposed an alternative to standard explanations of early phonetic learning in infants, as theories about this phenomenon rely heavily on the notion of phone categories.

With a similar method, Matusevych et al. (2020) tested the ability of various self-supervised speech models to reproduce infants' discrimination behaviour in multiple languages for a small set of

pairs of sounds. However, no quantitative comparison with behavioural data was made. Within the same test framework, Schatz and Feldman (2018) showed that a neural network trained to perform phone recognition was better at qualitatively reproducing Japanese and English native speakers' discrimination behaviour than an HMM-GMM model, focusing once again on the [r]/[l] pair of sound and also on vowel length differences. In this paper, we decide to: (i) evaluate different self-supervised speech models on more contrasts than these previous works (ii) directly compare their results with human behaviour (iii) measure models' similarity to humans at the *stimuli* level on top of doing it at the contrast level.

3 Methods

3.1 Human ABX test

Our probes of human speech perception use ABX phone discrimination tests, in which participants hear three speech extracts: A, B and X (an A/B/X **triplet**). A and B always differ in exactly one phone, and X is always (a distinct recording of) the same sequence of phones as either A or B (for example, A: /pap/, B: /pip/, X: /pap/). We ask the participants to indicate which of the first two sounds (A or B) is the most similar to the last sound (X). The ability of the participants to select the correct (*target*) rather than the distractor (*other*) speech extract indicates how well the population tested can discriminate the two phone categories p_1 and p_2 that *target* and *other* belong to (in our example, /i/ and /a/). We call p_1:p_2 a **contrast**. In this paper, we examine the results of monolingual French- and English-speaking participants.

3.2 Using models to predict

As in previous works (Millet et al., 2019; Millet and Dunbar, 2020a,b; Millet et al., 2021), to test models in the same way as participants, we extract a representation M for each of the three stimuli making up each A/B/X triplet in the experiment. We compute, for a triplet **target/other**/X, each model's Δ-value:

$$\Delta = DTW(M_{other}, M_X) - DTW(M_{target}, M_X) \quad (1)$$

with DTW being a distance obtained using dynamic time warping to aggregate a frame-level cosine distance along the warping path. The larger (more positive) the Δ-value obtained, the better the model is at discriminating the **target** and **other** phone categories. In our comparison between humans' and models' discrimination behaviour, we will generally use the raw Δ-values. The accuracy of the model on a specific triplet, independent of human listeners' behaviour, can also be computed by considering the model to be correct if the corresponding Δ value is greater than zero and incorrect otherwise. Below, we will refer to this objective accuracy as an **ABX score.**

3.3 Models

We compare self-supervised speech models to see if the representational spaces they develop during training on a language resemble humans' perceptual spaces. We choose to test three state-of-the-art self-supervised models: contrastive predictive coding (CPC), the basis for the current best-performing systems on the Zero Resource Speech Challenge evaluation (Dunbar et al., 2021); wav2vec 2.0; and a HuBERT model. These last two models obtain excellent word error rates on the task of semi-supervised speech recognition (self-supervised pre-training plus supervised fine-tuning on a small corpus).

As we use behavioural data from French and English-speaking participants, models are trained on either French or English recordings. To test for the impact of training on speech recordings compared to other types of sounds, we also train the models on recordings of acoustic scenes (non-speech). We choose one specific output layer for each model, using the one that obtains the best result in terms of human similarity.

We use classic acoustic features as a baseline, using the first 13 mel-frequency cepstrum coefficients (MFCCs), calculated using LIBROSA,[3] with a window of 25 ms and a stride of 10 ms. We also train DeepSpeech (Amodei et al., 2016) as a supervised reference.

3.3.1 Contrastive predictive coding

We use a light version of a model that uses contrastive predicting coding (CPC: Rivière et al. 2020). This model is smaller than HuBERT or wav2vec 2.0, as it is only made up of 5 convolutions (the encoder) and one LSTM layer (the sequence model). It is trained using a contrastive loss. For a sequential input $x = (x_1, ...x_t, ..., x_T)$, at time t, given the output of the sequential model, the loss

[3] https://librosa.org/

pushes the model to distinguish the K next outputs of the encoder in the future from randomly sampled outputs from another part of x. The detailed loss can be found in Appendix A. We use the output of the sequence model as representations for the CPC model.

3.3.2 Wav2vec 2.0

We test wav2vec 2.0 (Baevski et al., 2020). The model is made up of three elements: an encoder, a quantizer, and a decoder. The encoder is made up of five convolutional layers, the quantizer is a dictionary of possible representations, and the decoder is made up of 12 transformer layers. When an input z is given to the quantizer, it outputs the representation q from the dictionary that is the closest to the input. For an input x, wav2vec 2.0 uses the encoder to transform it into z, which is then quantized into q, and in parallel z is directly passed to the decoder to obtain a context representation c.

Like the CPC model, wav2vec 2.0 is trained using a contrastive loss L_m. Unlike the CPC model, it uses masking. Given a decoder representation of the context around some masked time step t, the loss pushes the model to identify the true quantized speech representation q_t from among a set of $K+1$ quantized candidate representations $\tilde{\mathbf{q}} \in \mathbf{Q}_t$ including q_t and K distractors uniformly sampled from other masked time steps in the same utterance (see Appendix A for details). We analyse the fifth layer of the decoder.

3.3.3 HuBERT

We also test a HuBERT model (Hsu et al., 2021a,b). This model uses exactly the same architecture as wav2vec 2.0 (except for the quantizer, which is not used), but with a different objective. Its training relies on an unsupervised teacher h (in our case, a K-means algorithm) that assigns a cluster label to each frame. Formally, we have $h(X) = Z = [z_1, ...z_T]$, with z_t a C-class categorical variable. HuBERT is trained to guess this cluster assignment for masked and unmasked frames at the same time. The detailed loss can be found in Appendix A.

The unsupervised teacher h is initially a K-means clustering on MFCCs. After a round of training using this initial teacher, h is replaced by a K-means model trained on the output of the sixth transformer layer of the model, and training restarts from scratch. We analyse the output of the sixth transformer layer.

3.3.4 Supervised reference: DeepSpeech

As a supervised reference system, we test a trained DeepSpeech model (Amodei et al., 2016). This model is not too intensive to train, is known to obtain reasonable ASR results, and has previously been compared to human speech perception (Millet and Dunbar, 2020b; Weerts et al., 2021). We train it to generate phonemic transcriptions.

DeepSpeech is composed of two convolutional layers followed by five RNN layers and a fully connected layer. The model is trained using spectrograms as input and a CTC loss, without a language model. We use representations extracted from the fourth RNN layer of the model, as it seems to give the best results, both in terms of absolute phone discriminability and for predicting human behaviour.

3.4 Comparing humans and models' perceptual space

In order to compare humans' and models' perceptual spaces, we use two metrics: the **log-likelihood** ($\ell\ell$) of a binary regression model on the experimental responses, and the **Spearman's ρ correlation** between the average of the model's Δ-values and participants' accuracies averaged within each phone contrast. These allow for predictions at two levels of granularity: the discriminability of individual experimental items ($\ell\ell$) and the overall discriminability of pairs of phones (ρ). In the default (**native**) setting, French-trained models are used to predict French-speaking participants' discrimination results, and similarly for English. See below for details.

For each model tested (see Section 3.3), we fit a probit regression to predict the binary responses of the participants (coded as correct or incorrect) using as a predictor the Δ values obtained from the model's representational space. In addition to a global intercept, the regression has other predictors to account for various nuisance factors: whether the right answer was A (1) or B (0); the order of the trial in the experimental list; a categorical predictor for the participant; and another for the Perceptimatic subset the result belongs to. We fit the model with an L1 regularisation (lasso). The $\ell\ell$ is obtained from the fitted regression model: the larger (less negative) the $\ell\ell$, the better the given model's Δ values predict the experimental data; thus, the more similar the model's representational space is to the perceptual space of the experimental participants.

We complement the log-likelihood metric with a correlation statistic. We compute the Spearman correlation (ρ), a correlation between the ranks of participants' accuracies (using their gradient results if available) and models' Δ-values, both averaged at the level of the phone contrast (zero indicates no correlation, one indicates a perfect monotonic relation). This measure averages out effects of individual A/B/X stimuli below the level of the phone contrast.

3.5 Comparing native language biases

Beyond global measures of how well models' representational spaces correspond to human listeners' perceptual spaces, we seek to assess how well the models reproduce group differences caused by the participants' native languages. One could think that humans are very good at discriminating all the sounds from their native language, and that they struggle to differentiate all the sounds from other languages. But reality is more complex than that: some contrasts are equally difficult or easy (even if they are not native) to discriminate for different language groups. The only way to study accurately native language biases is to focus on the relative discrimination difficulties shown by different language groups when listening to the same contrasts.

We present a method which evaluates the ability of the models to directly predict the relative difficulty of contrasts across the two language groups we have in the dataset we use. In other words, we measure if the models, when trained on French and English, show the same discrimination behaviour *differences* than French- and English-speaking participants.

We first normalise the Δ values obtained by each model by dividing by their standard deviation (within model/training condition, across all A/B/X triplets), in order to put the Δ values on the same scale for the two models. We average the normalised Δ values by contrast. We then calculate the overall accuracies for each phone contrast in the listening experiment.

We calculate difference scores: for each phone contrast, we subtract an English model's average Δ values from the average Δ value for the corresponding French-trained model. We do the same with the English-speaking and the French-speaking participants' contrast-level accuracy scores. This yields a measure of the *native language effect* for each phone contrast, for each model, and similarly for the human participants.

For each model, we compute a Pearson correlation between its contrast-level native language effects and those of human listeners. The closer the correlation is to one, the better the phone-level native language effects are captured by a given model.

Because this score calculates a native language effect independently for the models and for the participants, it is not susceptible to the same confounds as an approach which would derive the native language effect from a comparison of two different (and thus not necessarily comparable) models' fit to the data. Note, however, that the approach we propose is restricted to predicting contrast-level effects of native language.

4 Experiments

4.1 The Perceptimatic dataset

For the human data, we use five experiments from the Perceptimatic benchmark dataset,[4] containing the results of French- and English-speaking participants results on ABX phone discrimination experiments. Stimuli come from French, English, Brazilian Portuguese, Turkish, Estonian, and German, and test a variety of contrasts between vowel and consonant sounds, some of which are familiar, and some of which are unfamiliar, to the listeners. The five datasets use different kinds of stimulus triplets, including short three-phone extracts cut from running speech (**Zero Resource Speech Challenge 2017** and **Pilot July 2018** datasets), as well as read-speech nonwords, which highlight English consonants and vowels (**Pilot August 2018**), compare English with French vowels in a crosslinguistic task (**Cogsci-2019**), or highlight vowel contrasts in a variety of languages (**WorldVowels**). The combined dataset contains 4231 distinct triplets (each of which is sometimes presented to participants in the order target/other/X, sometimes in the order other/target/X), which test 662 phone contrasts, and contains data from 259 French-speaking participants and 280 English-speaking participants (not the same participants for all stimuli).

4.2 Models' training

The speech models we use are trained on 600-hour subsets of either the English or the French Com-

[4] See https://docs.cognitive-ml.fr/perceptimatic/ for access to, and more detailed descriptions of, the data.

monVoice datasets (Ardila et al., 2019). To train DeepSpeech as a phone recognizer, the text transcriptions included in CommonVoice are phonemized using eSpeakNG.[5] When English-trained models are used to predict English-speaking participants' results and French-trained for French-speaking participants', we refer to the trained models as **nat-cpc**, **nat-w2v**, **nat-hub**, and **nat-deep**.

To measure the impact of training on speech versus non-speech audio, the self-supervised models are also trained on a 595-hour subset of the Audioset dataset (Gemmeke et al., 2017) containing no human vocalizations.[6] We refer to these models as **aud-cpc**, **aud-w2v**, and **aud-hub**.

Each dataset is split randomly into train (80%), test (10%) and validation (10%). All recordings are resampled at 16000Hz and transformed into mono channel using sox.[7]

For the CPC model, we use the Facebook Research implementation[8] with all the default parameters. We train the model for 110 epochs and take the models that present the best loss on the validation set.

For wav2vec 2.0, we use the Fairseq Base implementation,[9] using the LibriSpeech configuration. As (Baevski et al., 2020), we train the models for 400k updates and take the model with the best loss on the validation set.

For HuBERT, we also use the Fairseq Base implementation[10] and the LibriSpeech configuration. We follow all the training settings of (Hsu et al., 2021a): our first-pass training takes its unsupervised teacher labels from a K-means algorithm with 50 clusters on the MFCCs for 10% of the training set, training for 250k updates. We then extract the representation of the training set from the sixth transformer layer and use these representations to train a new K-means with 100 clusters and re-train the model using these categories as the teacher for 450k updates. We use the model with the best loss on the validation set.

We use a PyTorch implementation of DeepSpeech.[11] We train the models for 150 epochs (to reach an overfitting point), saving a checkpoint of the model for each epoch. We then take the checkpoint that produces the best result in terms of Phone Error Rate (PER) on the validation set. We use specaugment (Park et al., 2019) to improve the model performance. The French model obtains 7.8% PER on the French test set and the English model obtains 22.75% PER on the English test set.

5 Results

In all graphs, statistical significance of comparisons is evaluated by bootstrapping over participants' results ($N = 10000$); redundant statistical comparisons are omitted for clarity (i.e. $C > A$ is omitted when $C > B$ and $B > A$). Confidence intervals shown are 95% bootstrap intervals.

5.1 Overall accuracy

Before using models' representational spaces to predict human discrimination behaviour, we look at how well models discriminate phones in their training language. We use the sign (positive/negative) of the Δ values to calculate the objective accuracy of selecting the target phone (**ABX scores**). For interpretability, we calculate scores only on the subsets of Perceptimatic containing monolingual English and French stimuli which were presented to listeners in their native language (**Zero Resource Speech Challenge 2017**, **WorldVowels**n and **Pilot August**). Results are shown in Table 1. In general, native self-supervised models obtain scores as good as or better than the supervised reference and human listeners, with a small preference for the **nat-w2v** model. They show a clear improvement over the corresponding models trained on acoustic scenes (non-speech). Certain datasets present more difficulties for the self-supervised models relative to **nat-deep**—notably, the English read-speech nonwords (from the **WorldVowels** and **Pilot August** subsets). Further details and comparison of ABX scores between native and non-native settings can be found in Appendix C.

5.2 Predicting human listeners

To assess how well self-supervised models' representational spaces match humans' perceptual spaces for speech, we compute the log-likelihood ($\ell\ell$) and the Spearman correlation (ρ) metrics over

[5] https://github.com/espeak-ng/espeak-ng

[6] A complete list of the labels kept can be found in our github: https://github.com/JAMJU/Sel_supervised_models_perception_biases

[7] http://sox.sourceforge.net/

[8] https://github.com/facebookresearch/CPC_audio

[9] https://github.com/pytorch/fairseq/tree/master/examples/wav2vec

[10] https://github.com/pytorch/fairseq/tree/master/examples/hubert

[11] https://github.com/SeanNaren/deepspeech.pytorch

	Zero		Vowels		PilotA
	Fr	En	Fr	En	En
Humans	0.84	0.80	0.80	0.84	0.74
MFCC	0.76	0.77	0.73	0.76	0.88
nat-deep	0.82	0.83	0.75	**0.87**	**0.94**
nat-cpc	0.85	0.85	0.67	0.83	0.85
aud-cpc	0.76	0.74	0.55	0.72	0.66
nat-w2v	**0.88**	**0.88**	0.71	0.83	0.84
aud-w2v	0.76	0.73	0.53	0.71	0.78
nat-hub	0.87	0.87	**0.76**	0.83	0.82
aud-hub	0.77	0.78	0.57	0.77	0.74

Table 1: ABX scores on three subsets of the Perceptimatic dataset, each containing a French and an English subset; the larger (closer to one) the better. Scores are averages over the per-triplet accuracies. Models are native-language models, except those trained on AudioSet. Bold scores are the best in the column.

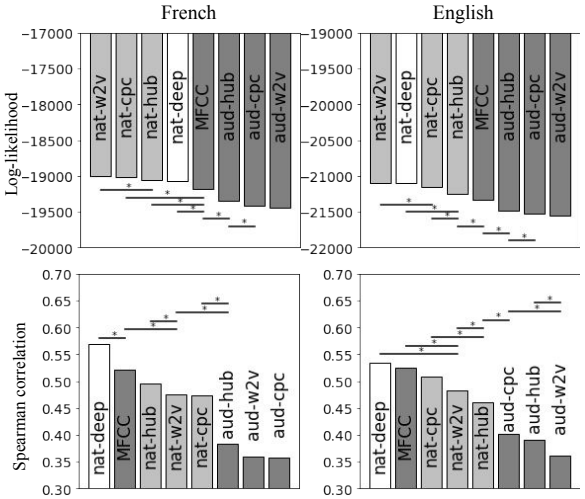

Figure 1: Log-likelihood values (top: shorter/higher bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the native self-supervised models in light grey and the baselines in darker grey (neutral acoustic features and models trained on acoustic scenes).

the entire Perceptimatic dataset (see Section 3.4) in the native-language training condition. Results can be seen in Figure 1.

First, we need to note that the models' performance appears to be importantly tied to training on speech, rather than simply on natural audio. Indeed, the models trained on acoustic scenes (non-speech) consistently perform worse than the native-trained models and MFCCs, on both measures.

For the $\ell\ell$ metric, **nat-w2v** does at least as well as, or (for French) somewhat better than, the supervised reference at modelling human listeners' perceptual confusions; most native self-supervised models perform similarly. Self-supervised models appear to learn representational spaces at least as similar to human native listeners' as our supervised phone recogniser when measured in this way.

The ρ metric, which correlates models' with humans' average dissimilarity (Δ or accuracy) for each phone contrast, reveals a different pattern. Here, **nat-deep** performs best. Furthermore, native self-supervised models perform worse than generic MFCC features. This suggests a component of human speech perception that is poorly captured by self-supervised models at the contrast level. (On some subsets—notably the **WorldVowels** set of familiar and unfamiliar vowel contrasts—self-supervised models *are* better than MFCCs, but are still worse than our supervised reference; see Appendix B.)

To confirm the difference of result for the contrast level (the ρ metric) and the stimuli level (the $\ell\ell$ metric), we compute the Spearman correlation metric at the stimuli level, averaging participants' results over the stimuli, instead of doing it, for models and humans, over contrasts. The results of this analysis can be found in Figure 2. We notice that this new analysis, done at the stimuli level, gives similar results than our log-likelihood metric. This supports the idea that the bad results for the original ρ metric of the self-supervised models we consider are due to the averaging over contrast.

To illustrate the comparisons at the level of phone contrasts, in Figure 3 we plot the average accuracy (per contrast) for French-speaking participants results against (left) DeepSpeech trained on French, one of the best-performing models, and (right) wav2vec 2.0 trained on AudioSet (aud-w2v), one of the models that is the least similar to humans.

5.3 Native language biases

To look for the presence of human-like native language biases, we look at the ability of native models to predict the difference in behaviour between the French- and the English-speaking groups (see Section 3.5). Figure 4 (left) shows the native language effect assessed over the entire Perceptimatic dataset—that is, the correlation, at the contrast

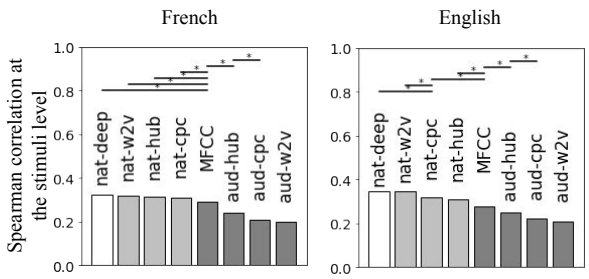

Figure 2: Spearman correlation at the stimuli level (taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant.

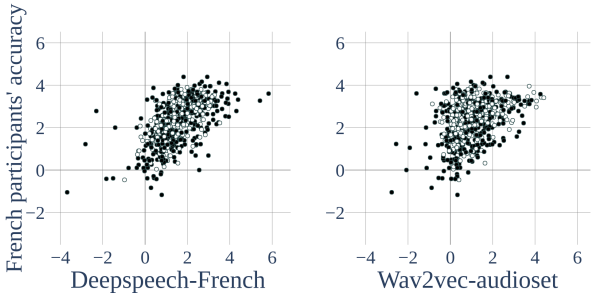

Figure 3: Average of French listeners' results (higher: better discrimination) against average δ from (**left**) supervised reference trained on phonemic transcriptions (**right**) wav2vec trained on non-speech recordings. Each point is a contrast. Measures are normalised by dividing by standard deviation over the entire data set, so the two scales are comparable. Black circles are non-native contrasts, white ones are native (French).

level, between the differences in Δ across language-training conditions, on the one hand, and the differences in accuracy for the two listener groups, on the other. **Nat-cpc** is competitive with **nat-deep** at predicting differences in groups. **Nat-hub** and **nat-w2v**, on the other hand, show very native language effect.

Figure 4 (right) shows the same analysis, but on only the **WorldVowels** dataset. The stimuli in this dataset are constructed to specifically induce different discrimination behaviour between the two language groups. Here, **nat-deep** shows a much better ability to predict native language effects, both in the absolute, and relative to the other models.

As this analysis is done at the level of phone contrasts, and not individual stimuli, we could think that as our supervised reference model is trained to produce phonemic transcriptions, it probably gives it a head start at predicting differences in discrimination behaviour driven by phone categories. To look more precisely at this, we compute our native effect at the stimuli level instead of the contrast level. The results of this analysis can be seen in Figure 5, for the all dataset and for the WorldVowels subset. Going to the stimuli level reduces radically the native effect measured. This is expected, as the number of participants' result per stimulus is small, and the effect measured on humans is thus very noisy when measured at this level, and therefore harder to reproduce for the models. However, we can notice that our supervised reference and the CPC model are still the ones that exhibit the most native language effect.

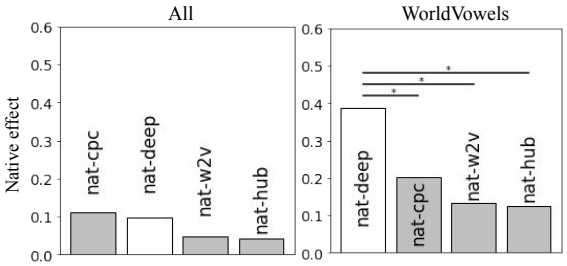

Figure 4: Native language effect for each model, the bigger the bar, the better the models capture language specificities in the discrimination behaviour between the two groups. Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the self-supervised models in light grey.

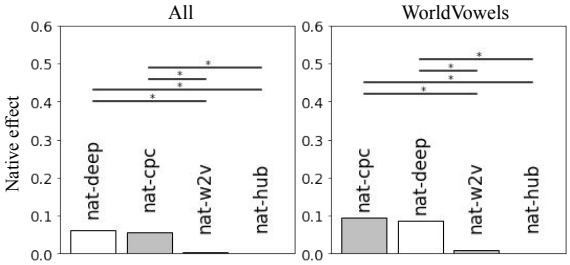

Figure 5: Native language effect for each model, the bigger the bar, the better the models capture language specificities in the discrimination behaviour between the two groups. Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the self-supervised models in light grey.

6 Discussion

We showed that the self-supervised models we tested seem to learn representational spaces relevant for predicting human phone discrimination at the stimuli level. However, while humans show

consistent discrimination behaviour for certain contrasts, whatever the stimuli, the self-supervised models we test do not capture systematic effects of contrasts between specific pairs of phones. Unlike our supervised reference, their similarity to human perceptual spaces is limited to capturing the discriminability of specific individual stimuli. The models tested were similar, but wav2vec 2.0 showed a slight advantage for predicting this kind of behaviour.

We have also shown that training on speech data is essential to obtaining a human-like perceptual space: for all of our metrics (ABX accuracy or similarity to humans), training on speech leads to better results than training on acoustic scenes (non-speech). This strongly suggests that the benefits of self-supervised speech models comes from learning characteristics of human speech, not simply the fact that they are better general audio features. We speculate that this is not just important to their ability to predict human speech perception and to discriminate phones, but also of their (related) utility for doing downstream tasks such as ASR.

What these models learn about speech, however, is not typically language-specific—at least, not in the same way that human perception is. Wav2vec 2.0 and HuBERT do not model language-specific differences in human speech perception, and can be seen as modelling a language-neutral or universal speech perception space. Indeed, they exhibit very few native language effect (see Figure 4 and 5). We note that the idea of self-supervised models learning universal speech features is consistent with the fact that models trained on one language, or multilingually, have proven useful for representing speech in unseen languages (Riviere et al., 2020).

CPC does capture effects of native language on perception at the contrast level, but to a far lesser extent than our supervised reference when we focus on a subset of Perceptimatic designed to capture important differences in discrimination behaviour for our two groups of participants (WorldVowels). Our CPC model differs from the other models tested in its small size, its causal architecture (wav2vec and HuBERT use transformers), and in that it does not use masking during its training. Its architecture is probably the most biologically plausible of the three self-supervised models we tested. We should note, however, that it does not make it the best predictor of human discrimination behaviour among the three models (see Figure 1 and 2).

One possible explanation for the self-supervised models' limitations we observe is insufficiency of training data: the models in question have generally shown good performance on downstream tasks when pre-trained on large amounts of data. We tested this using available pretrained wav2vec and HuBERT models trained on much larger amounts of data. The detailed results can be found in Appendix E. The models show a slight improvement, but, when looking at the ρ statistic at the phone contrast level, they are still worse than MFCCs.

Contrary to previous results (Millet and Dunbar, 2020a,b), our supervised reference system is quite good at predicting human discrimination behaviour (in particular at the contrast level), and clearly predicts a native language effect. The main differences in our experiment with (Millet and Dunbar, 2020b) are the type of model (DeepSpeech instead of HMM-GMM), and with (Millet and Dunbar, 2020a) the type of training objective (phone recognition rather than prediction of orthographic text), and the size of the training corpora (we use fewer data). Predicting phones rather than orthography seems to be critical (as we demonstrate in Appendix F), and using a neural network instead of a Bayesian model (HMM-GMM) leads to a more human-like representational space, as already highlighted by (Schatz and Feldman, 2018).

Given the advantage supervised phone recognizers show, a different approach to developing more human-like representational spaces in self-supervised models might be the inclusion of tasks or constraints that push them to take into account longer time scales in order to encourage them to construct longer, more phone-like units.

Acknowledgements

This research was supported by the École Doctorale Frontières du Vivant (FdV) – Programme Bettencourt, by the Connaught Fund and the Arts and Science Tri-Council Bridging Fund, University of Toronto, and by French Agence Nationale de la Recherche grants ANR-17-CE28-0009 (GEOMPHON), ANR-11-IDFI-023 (IIFR), ANR-11-IDEX-0005 (USPC), ANR-10-LABX-0083 (EFL), ANR-17-EURE-0017 Frontcog, ANR-10-IDEX-0001-02 PSL*, ANR-19-P3IA-0001 PRAIRIE 3IA Institute. This work was performed using HPC resources from GENCI-IDRIS (Grant 2021-AD011012415, 2022-AD011012415R1).

References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

MP Cooke and OE Scharenborg. 2008. The interspeech 2008 consonant challenge.

Ewan Dunbar, Mathieu Bernard, Nicolas Hamilakis, Tu Nguyen, Maureen de Seyssel, Patricia Rozé, Morgane Rivière, Eugene Kharitonov, and Emmanuel Dupoux. 2021. The zero resource speech challenge 2021: Spoken language modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1.

Jenelle Feather, Alex Durango, Ray Gonzalez, and Josh McDermott. 2019. Metamers of neural networks reveal divergence from human perceptual systems. In *NeurIPS*, pages 10078–10089.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780. IEEE.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021a. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021b. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE.

Patricia K Kuhl, Barbara T Conboy, Denise Padden, Tobey Nelson, and Jessica Pruitt. 2005. Early speech perception and later language development: Implications for the" critical period". *Language learning and development*, 1(3-4):237–264.

Patricia K Kuhl, Erica Stevens, Akiko Hayashi, Toshisada Deguchi, Shigeru Kiritani, and Paul Iverson. 2006. Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science*, 9(2):F13–F21.

Kushal Lakhotia, Evgeny Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh Nguyen, Jade Copet, Alexei Baevski, Adelrahman Mohamed, et al. 2021. Generative spoken language modeling from raw audio. *arXiv preprint arXiv:2102.01192*.

Erika S Levy. 2009. On the assimilation-discrimination relationship in american english adults' french vowel learning. *The Journal of the Acoustical Society of America*, 126(5):2670–2682.

Yevgen Matusevych, Thomas Schatz, Herman Kamper, Naomi H Feldman, and Sharon Goldwater. 2020. Evaluating computational models of infant phonetic learning across languages. *arXiv preprint arXiv:2008.02888*.

Bernd T Meyer, Tim Jürgens, Thorsten Wesker, Thomas Brand, and Birger Kollmeier. 2010. Human phoneme recognition depending on speech-intrinsic variability. *The Journal of the Acoustical Society of America*, 128(5):3126–3141.

Juliette Millet, Ioana Chitoran, and Ewan Dunbar. 2021. Predicting non-native speech perception using the perceptual assimilation model and state-of-the-art acoustic models. In *CoNLL 2021 Proceedings, 25th Conference on Computational Natural Language Learning*.

Juliette Millet and Ewan Dunbar. 2020a. Perceptimatic: A human speech perception benchmark for unsupervised subword modelling. *2020 Interspeech Conference Proceedings*.

Juliette Millet and Ewan Dunbar. 2020b. The perceptimatic english benchmark for speech perception models. *2020 Cogsci Conference Proceedings*.

Juliette Millet, Nika Jurov, and Ewan Dunbar. 2019. Comparing unsupervised speech learning directly to human performance in speech perception. In *CogSci 2019-41st Annual Meeting of Cognitive Science Society*.

Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation

method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Morgane Rivière and Emmanuel Dupoux. 2021. Towards unsupervised learning of speech features in the wild. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 156–163. IEEE.

Morgane Riviere, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7414–7418. IEEE.

Morgane Rivière, Armand Joulin, Pierre-Emmanuel Mazaré, and Emmanuel Dupoux. 2020. Unsupervised pretraining transfers well across languages.

Odette Scharenborg, Sebastian Tiesmeyer, Mark Hasegawa-Johnson, and Najim Dehak. 2018. Visualizing phoneme category adaptation in deep neural networks. In *Interspeech*, pages 1482–1486.

Thomas Schatz. 2016. *ABX-discriminability measures and applications*. Ph.D. thesis, Université Paris 6 (UPMC).

Thomas Schatz and Naomi H Feldman. 2018. Neural network vs. hmm speech recognition systems as models of human cross-linguistic phonetic perception. In *Proceedings of the conference on cognitive computational neuroscience*.

Thomas Schatz, Naomi H Feldman, Sharon Goldwater, Xuan-Nga Cao, and Emmanuel Dupoux. 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7).

Feng-Ming Tsao, Huei-Mei Liu, and Patricia K Kuhl. 2004. Speech perception in infancy predicts language development in the second year of life: A longitudinal study. *Child development*, 75(4):1067–1084.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Lotte Weerts, Stuart Rosen, Claudia Clopath, and Dan FM Goodman. 2021. The psychometrics of automatic speech recognition. *bioRxiv*.

Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. 2021. Self-training and pre-training are complementary for speech recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3030–3034. IEEE.

Reiko A Yamada and Yoh'ichi Tohkura. 1990. Perception and production of syllable-initial english/r/and/l/by native speakers of japanese. In *First international conference on spoken language processing*.

Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.

A Detailed losses used by the models

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

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

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:

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

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.

B 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 7, for WorldVowels in Figure 6, for Zerospeech in Figure 10, for pilot-july in Figure 8, and for pilot-august in Figure 9. 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.

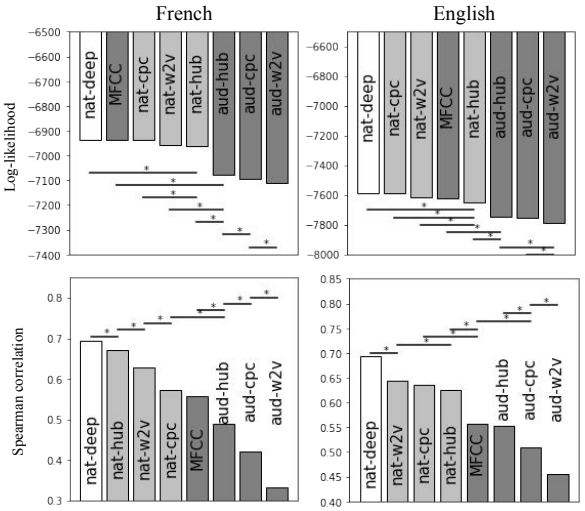

Figure 6: Results on the **WorldVowels** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference 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).

C Difference in ABX score between French and English models

To complete Table 1, we present in Figure 11 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).

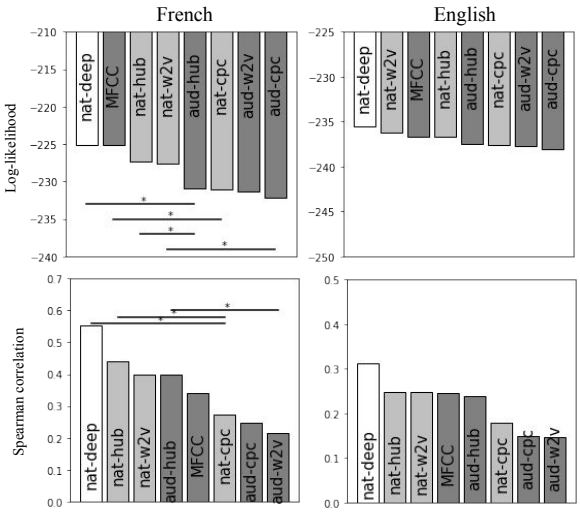

Figure 7: Results on the **Cogsci-2019** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference 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).

D Language preference

A possible approach to study models' language specificity would be to see if English-trained models predict English-speaking participants better than French-trained models, and vice versa. We assess whether models in the native training condition predict discriminability better than the corresponding models in the non-native training condition. Figure 12 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

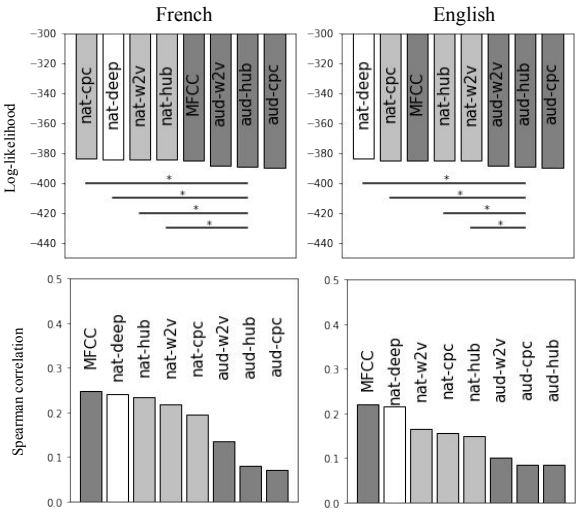

Figure 8: Results on the **pilot-july-2018** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference 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).

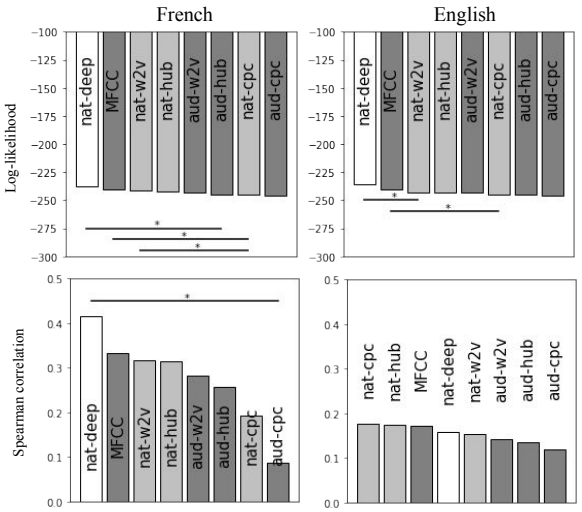

Figure 9: Results on the **pilot-august-2018** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference 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).

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.

E 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 13. Their different ABX scores can also be seen in Table 2. Models trained on English are evaluated on English-

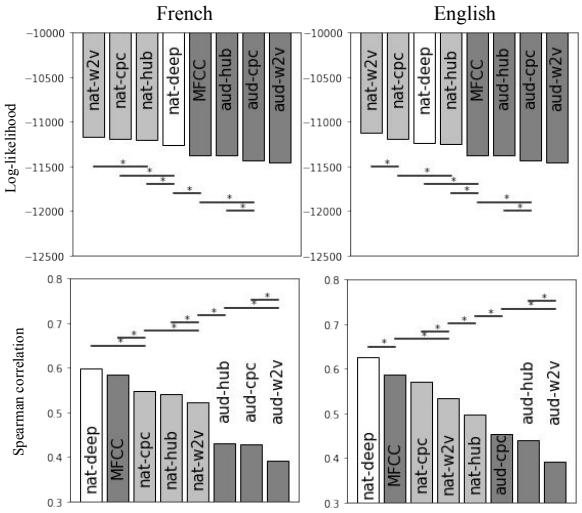

Figure 10: Results on the **Zerospeech** subset. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference 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).

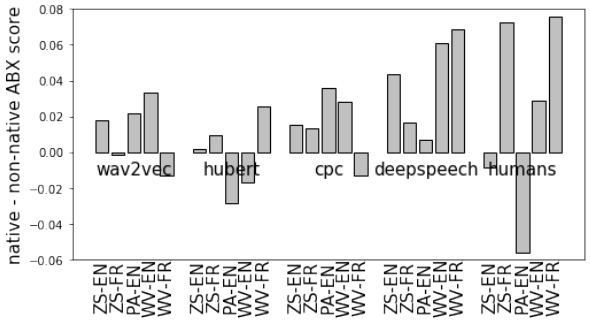

Figure 11: ABX score difference between native setting and non-native setting for the different models tested. The bigger the bar above zero, the bigger difference.

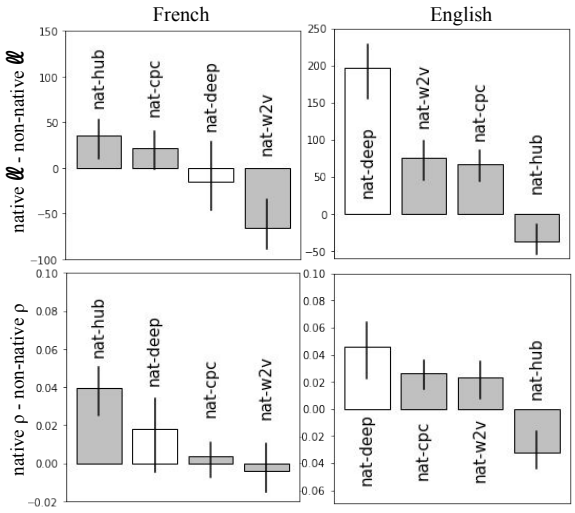

Figure 12: Native minus non-native log-likelihood values (top) and Spearman correlations (bottom) for French (*left*) and English participants (*right*). The higher the bar above zero, the better the native setting is compared to the non-native setting. The supervised reference is in white, the self-supervised models are in light grey. Black lines indicate 95% confidence intervals.

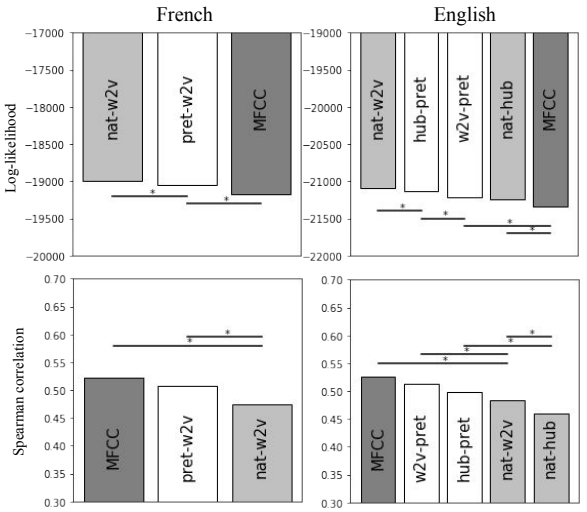

Figure 13: Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The pretrained models are in white to distinguish it from our self-supervised models trained on only 600h of speech.

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

speaking participants (and English contrast for the ABX scores), and same for French.

F 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 14. A comparison of English-speaking participants' discrimination ability and the two supervised models' Δ-values can also be seen in Figure 15. Models trained on phonemic transcriptions are better at

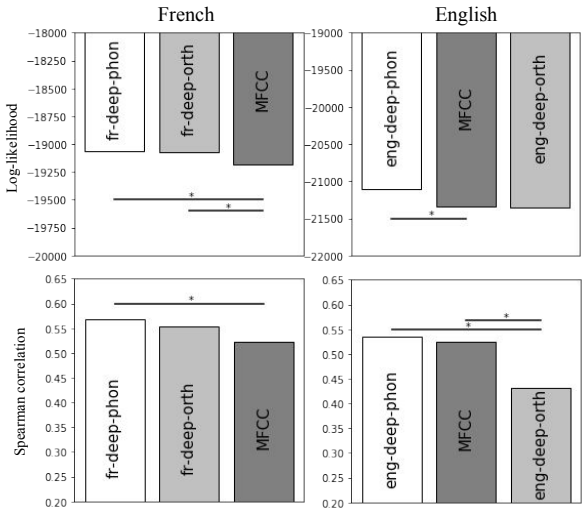

Figure 14: Results of DeepSpeech trained on phonemic transcriptions (phon) or orthographic (orth), compared with MFCCs. Log-likelihood values (top: shorter bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant.

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. Appendix E 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.

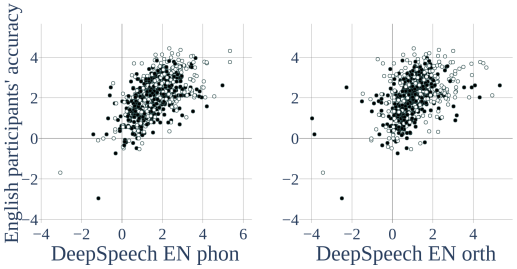

Figure 15: Average of English listeners' results (higher: better discrimination) against average δ from (**left**) supervised reference trained on phonemic transcriptions (**right**) trained on orthographic transcriptions. Each point is a contrast. Measures are normalized by dividing by standard deviation over the entire data set. Black circles are non-native contrasts, white ones are native (English).