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

Juliette Millet,[1 2 3] **Ewan Dunbar,** [1 4]

[1] CoML, ENS/CNRS/EHESS/INRIA/PSL, Paris, France
[2] LLF, University of Paris, CNRS, Paris, France
[3] CRI, FAN, IIFR, University of Paris, Paris, France
[4] University of Toronto, Toronto, Canada
juliette.millet@cri-paris.org, 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, 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 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, and effects of listeners' native language, on perception.

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, Vinyals, and Kavukcuoglu 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) (***right*** and ***light*** 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 ***rue*** (/y/: "street") and ***roue*** (/u/: "wheel"). 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 dissimilarities. 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). Unlike in previous works, we do not perform supervised fine-tuning on wav2vec 2.0. 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

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

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 the learned features are better than generic acoustic features, not only for applied tasks like phone recognition, but also for modelling effects of acoustic similarity on human speech perception. However, the models' representational spaces, while similar to those of humans, miss two critical properties: categorical effects of phones, and a strong influence of native (training) language.

All our code and data are freely available.[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 showed 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, Jurov, and Dunbar 2019; Millet and Dunbar 2020a,b; Millet, Chitoran, and Dunbar 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 on 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, which 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 reproducing human discrimination behaviour than an HMM-GMM model, focusing once again on the [r]/[l] pair of sound for Japanese and English native speakers, and on vowel length differences which are informative for the former but not for the latter.

Methods

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.

Using models to predict

As in previous works (Millet, Jurov, and Dunbar 2019; Millet and Dunbar 2020a,b; Millet, Chitoran, and Dunbar 2021), to test models in the same way as participants, we extract the 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.**

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

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

HuBERT model. These last two obtain excellent word error rates on the task of semi-supervised speech recognition (self-supervised pretraining plus supervised fine-tuning on a small corpus). Models are trained on either French or English recordings, and on recordings of acoustic scenes (non-speech). 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.

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 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. We use the output of the sequence model as representations for the CPC model.

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 for details). We analyse the fifth layer of the decoder.

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.

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.

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.

Comparing humans and models' perceptual space

In order to compare humans' and models' perceptual spaces, we use two metrics (Millet, Chitoran, and Dunbar 2021): 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 **Models**), 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 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.

Comparing native language biases

Beyond global measures of how well models' representational spaces correspond to human listeners' perceptual

[3] https://librosa.org/

spaces, we seek to assess how well the models reproduce group differences caused by the participants' native languages.

A very simple way to study the impact of the training language on models' representational spaces would assess objectively whether French-trained models discriminate French-language A/B/X triplets better than do English models, and vice versa. However, not all native-language contrasts are created equal: some French contrasts are more difficult than others, even for French-speaking participants. Furthermore, this limits our evaluation to native-language stimuli.

Another possible approach would be to see if English-trained models predict English-speaking participants better than French-trained models, and vice versa. However, a problem arises with this method for comparing models if training a specific model on a specific language results in predictions that are globally better than other training configurations—that is, better at predicting *both* French- and English-speaking participants. Thus, this analysis can be hard to interpret. We show results in the appendix, but we do not present them in the main text for the sake of clarity.

To address this interpretation problem, we present a method which uses the models to directly predict the relative difficulty of contrasts across the two participant groups. 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 phone-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 phone-level effects of native language.

Experiments

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).

Model training

The models introduced above are trained using 600-hour subsets of the English and the French CommonVoice 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 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. We train the

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

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

[6] A complete list of the labels kept can be found in the [anonymous for review] Github repository for the paper.

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

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

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

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. Following (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 and take the checkpoint that produces the best result in term 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.

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.

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,** 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 wav2vec 2.0 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 DeepSpeech—notably, the English read-speech nonwords (from the **WorldVowels** and **Pilot August** subsets). Further details can be found in the appendix.

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 the entire Perceptimatic dataset (see **Comparing humans and models' perceptual space**) in the native-language training condition. Results can be seen in Figure 1.

	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
DpSpeech	0.82	0.83	0.75	**0.87**	**0.94**
CPC	0.85	0.85	0.67	0.83	0.85
AudioSet	0.76	0.74	0.55	0.72	0.66
wav2vec	**0.88**	**0.88**	0.71	0.83	0.84
AudioSet	0.76	0.73	0.53	0.71	0.78
HuBERT	0.87	0.87	**0.76**	0.83	0.82
AudioSet	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.

For the $\ell\ell$ metric, wav2vec 2.0 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 radically different pattern. Here, DeepSpeech 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. (On some subsets—notably the **WorldVowels** set of familiar and unfamiliar vowel contrasts—self-supervised models *are* better than MFCCs; see Appendix.)

The models' performance appears to be importantly tied to training on speech, rather than simply on natural audio. The models trained on non-speech consistently perform worse than the native-trained models, and than MFCCs, on both measures.

To illustrate the comparisons at the level of phone contrasts, in Figure 2 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, one of the worst-performing models.

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 **Comparing native language biases**). Figure 3 (left) shows the native language effect assessed over the entire Perceptimatic dataset—that is, the correlation, at the contrast level, between the differences in Δ

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

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

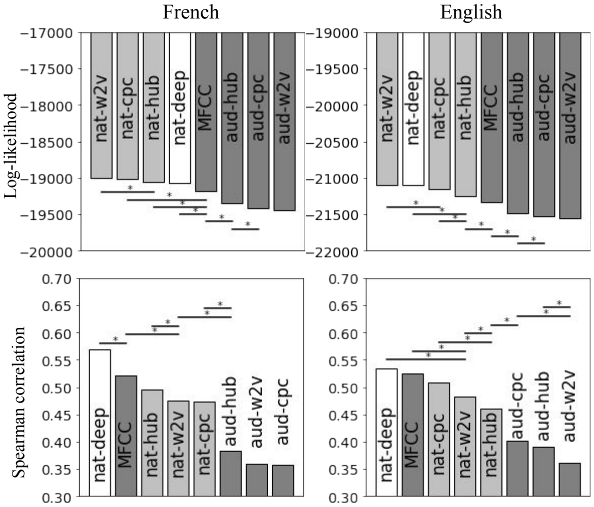

Figure 1: Log-likelihood values (top: shorter/higher bars are better) and Spearman correlation (bottom: taller bars are better) for French (*left*) and English participants (*right*). Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from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

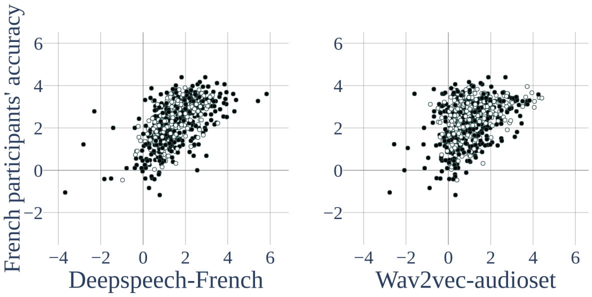

Figure 2: Average of French listeners' results (higher: better discrimination) against average δ from (**left**) supervised reference trained on phonemic transcriptions (**right**) wav2vec trained on non-speech recordings. Each point is a contrast. Measures are normalised by dividing by standard deviation over the entire data set, so the two scales are comparable. Black circles are non-native contrasts, white ones are native (French).

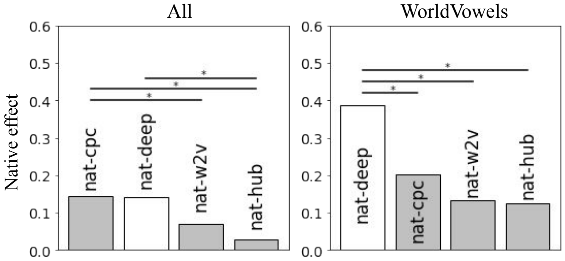

Figure 3: Native language effect for each model, the bigger the bar, the better the models capture language specificities in the discrimination behaviour between the two groups. Stars indicate that the pairwise difference is significant. The supervised reference is in white to distinguish it from the self-supervised models in light grey.

across language-training conditions, on the one hand, and the differences in accuracy for the two listener groups, on the other. As before, DeepSpeech performs well; unlike some of the previous results at the phone contrast level, CPC is competitive with DeepSpeech at predicting differences in groups. HuBERT and wav2vec 2.0, on the other hand, appear to be learning a language-neutral speech perception space.

Figure 3 (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, DeepSpeech 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, the fact that our supervised reference model is trained to produce phonemic transcriptions probably gives it a head start at predicting differences in discrimination behaviour driven by phone categories.

Discussion

We showed that the self-supervised models we tested seem to learn representational spaces relevant for predicting human phone discrimination. However, while human speech perception is known to be influenced by the set of phones in the native phone inventory—which renders many unfamiliar phones systematically difficult to distinguish—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, training on speech leads to better results than training on acoustic scenes. 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. 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, but to a far lesser extent than our supervised reference. 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.

One possible explanation for the 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 the appendix. 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, and clearly predicts a native language effect. The main differences in our experiment are the type of model (DeepSpeech), the type of training objective (phone recognition rather than prediction of orthographic text), and the size of the training corpora (we use less data). Predicting phones rather than orthography seems to be critical (as we demonstrate in the appendix).

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 20XX-AD011012415).

References

Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, 173–182. PMLR.

Ardila, R.; Branson, M.; Davis, K.; Henretty, M.; Kohler, M.; Meyer, J.; Morais, R.; Saunders, L.; Tyers, F. M.; and Weber, G. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Baevski, A.; Zhou, H.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *arXiv preprint arXiv:2006.11477*.

Cooke, M.; and Scharenborg, O. 2008. The interspeech 2008 consonant challenge.

Dunbar, E.; Bernard, M.; Hamilakis, N.; Nguyen, T.; de Seyssel, M.; Rozé, P.; Rivière, M.; Kharitonov, E.; and Dupoux, E. 2021. The Zero Resource Speech Challenge 2021: Spoken language modelling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–1.

Feather, J.; Durango, A.; Gonzalez, R.; and McDermott, J. 2019. Metamers of neural networks reveal divergence from human perceptual systems. In *NeurIPS*, 10078–10089.

Gemmeke, J. F.; Ellis, D. P.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R. C.; Plakal, M.; and Ritter, M. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 776–780. IEEE.

Hsu, W.-N.; Bolte, B.; Tsai, Y.-H. H.; Lakhotia, K.; Salakhutdinov, R.; and Mohamed, A. 2021a. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. *arXiv preprint arXiv:2106.07447*.

Hsu, W.-N.; Tsai, Y.-H. H.; Bolte, B.; Salakhutdinov, R.; and Mohamed, A. 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)*, 6533–6537. IEEE.

Lakhotia, K.; Kharitonov, E.; Hsu, W.-N.; Adi, Y.; Polyak, A.; Bolte, B.; Nguyen, T.-A.; Copet, J.; Baevski, A.; Mohamed, A.; et al. 2021. Generative spoken language modeling from raw audio. *arXiv preprint arXiv:2102.01192*.

Levy, E. S. 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.

Matusevych, Y.; Schatz, T.; Kamper, H.; Feldman, N. H.; and Goldwater, S. 2020. Evaluating computational models of infant phonetic learning across languages. *arXiv preprint arXiv:2008.02888*.

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

Millet, J.; Chitoran, I.; and Dunbar, E. 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*.

Millet, J.; and Dunbar, E. 2020a. Perceptimatic: A human speech perception benchmark for unsupervised subword modelling. *2020 Interspeech Conference Proceedings*.

Millet, J.; and Dunbar, E. 2020b. The Perceptimatic English Benchmark for Speech Perception Models. *2020 Cogsci Conference Proceedings*.

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

Oord, A. v. d.; Vinyals, O.; and Kavukcuoglu, K. 2017. Neural discrete representation learning. *arXiv preprint arXiv:1711.00937*.

Panayotov, V.; Chen, G.; Povey, D.; and Khudanpur, S. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 5206–5210. IEEE.

Park, D. S.; Chan, W.; Zhang, Y.; Chiu, C.-C.; Zoph, B.; Cubuk, E. D.; and Le, Q. V. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

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

Riviere, M.; Joulin, A.; Mazaré, P.-E.; and Dupoux, E. 2020. Unsupervised pretraining transfers well across languages. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 7414–7418. IEEE.

Rivière, M.; Joulin, A.; Mazaré, P.-E.; and Dupoux, E. 2020. Unsupervised pretraining transfers well across languages. arXiv:2002.02848.

Scharenborg, O.; Tiesmeyer, S.; Hasegawa-Johnson, M.; and Dehak, N. 2018. Visualizing Phoneme Category Adaptation in Deep Neural Networks. In *Interspeech*, 1482–1486.

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

Schatz, T.; and Feldman, N. H. 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*.

Schatz, T.; Feldman, N. H.; Goldwater, S.; Cao, X.-N.; and Dupoux, E. 2021. Early phonetic learning without phonetic categories: Insights from large-scale simulations on realistic input. *Proceedings of the National Academy of Sciences*, 118(7).

Wang, C.; Riviere, M.; Lee, A.; Wu, A.; Talnikar, C.; Haziza, D.; Williamson, M.; Pino, J.; and Dupoux, E. 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)*, 993–1003. Online: Association for Computational Linguistics.

Weerts, L.; Rosen, S.; Clopath, C.; and Goodman, D. F. 2021. The Psychometrics of Automatic Speech Recognition. *bioRxiv*.

Xu, Q.; Baevski, A.; Likhomanenko, T.; Tomasello, P.; Conneau, A.; Collobert, R.; Synnaeve, G.; and Auli, M. 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)*, 3030–3034. IEEE.

Yamada, R. A.; and Tohkura, Y. 1990. Perception and production of syllable-initial English/r/and/l/by native speakers of Japanese. In *First international conference on spoken language processing*.

Zhang, Y.; Qin, J.; Park, D. S.; Han, W.; Chiu, C.-C.; Pang, R.; Le, Q. V.; and Wu, Y. 2020. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*.