The Perceptimatic English Benchmark for Speech Perception Models

Juliette Millet, Ewan Dunbar

The Perceptimatic English Benchmark for Speech Perception Models

Juliette Millet (juliette.millet@cri-paris.org)
Université de Paris, LLF, CNRS, Paris, France
CoML, ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France
CRI, Département Frontières du Vivant et de l'Apprendre, IIFR, Université de Paris

Ewan Dunbar (ewan.dunbar@univ-paris-diderot.fr)
Université de Paris, LLF, CNRS, Paris, France
CoML, ENS/CNRS/EHESS/INRIA/PSL Research University, Paris, France

Abstract

We present the Perceptimatic English Benchmark, an open experimental benchmark for evaluating quantitative models of speech perception in English. The benchmark consists of ABX stimuli along with the responses of 91 American English-speaking listeners. The stimuli test discrimination of a large number of English and French phonemic contrasts. They are extracted directly from corpora of read speech, making them appropriate for evaluating statistical acoustic models (such as those used in automatic speech recognition) trained on typical speech data sets. We show that phone discrimination is correlated with several types of models, and give recommendations for researchers seeking easily calculated norms of acoustic distance on experimental stimuli. We show that DeepSpeech, a standard English speech recognizer, is more specialized on English phoneme discrimination than English listeners, and is poorly correlated with their behaviour, even though it yields a low error on the decision task given to humans.

Keywords: benchmarks; speech perception; acoustic distance; speech recognition

Introduction

There is no accurate computational model of human speech perception that applies to real speech. Implemented speech perception models exist which take artificial phonetic or perceptual features as input and map them to recognized words (McClelland & Elman, 1986; Norris & McQueen, 2008), use speech recognizers as a front-end to derive phonetic transcriptions (Scharenborg, Norris, ten Bosch, & McQueen, 2005), or work on raw speech waveforms for extremely artificial utterances only (Elman & McClelland, 2015). Yet, traditional automatic speech recognition systems directly analyze natural, recorded, continuous speech and decode it as a sequence of phonemes or words. We take the *reverse engineering* approach (Dupoux, 2018) of concluding that the signal processing and machine learning tools underlying automatic speech recognition should thus provide a starting point for a model of human speech perception.

Little is known, however, about the exact nature of the difference between the behaviour of human beings and that of speech processing tools developed for an applied purpose. We propose the **Perceptimatic English Benchmark** (PEB), an experimental human data set documenting English and French phone discrimination by English speaking individuals, which is amenable to comparisons with a wide range of models.[1]

Our ultimate goal is to build models of human speech perception. The narrower project to which this paper contributes is to build a testing architecture, and in particular to construct a series of different benchmarks, tapping into different aspects of human speech perception. Here, we construct an experiment which taps into a relatively detail-oriented mode of speech perception, constraining it to be as similar as possible to an existing data set (already used for testing statistical acoustic models) so as to be useful in applied unsupervised speech recognition research as well. We focus on a simple experiment for which typical speech recognition models could in principle give results comparable to humans, that of phone discrimination (typical speech recognition models are classifiers for sequences of phones). However, speech recognition models are trained on databases of continuous, natural speech, while typical experimental stimuli are individual phones, syllables, or words, read or synthesized in an effort to ensure that the phonetic properties being probed are audible. Such word-list type pronunciations, while clear to human listeners, are likely quite different from the training data of standard speech recognition models. Models applied to them would be faced with the often difficult task of generalizing to a novel speech style. We thus start with a conservative test: the Perceptimatic English Benchmark is constructed out of snippets from a French and an English corpus of read speech—*ecological* for typical models—tested as phone discrimination experiment items on English listeners.

To the degree made possible by the speech corpora from which the stimuli are extracted, we make the evaluation *complete,* in the sense that it tests discrimination of as many pairs of phones as possible, while being *controlled* in several ways, notably in never comparing phones extracted from radically different phonetic contexts. Using French stimuli in addition to English stimuli enables us to test English speaking individuals and models trained on English recordings on unfamiliar sounds, and study their foreign language speech perception. Details are found in **Perceptimatic English Benchmark** below.

In this paper, we use the PEB to evaluate seven models that apply to real speech. We compare models' representational space with human perceptual space by studying how

[1] All stimuli, human experimental data, analysis and processing scripts, and model results, are available at the following permanent link: https://github.com/JAMJU/Cogsci2020-Perceptimatic-English

well distances in models' representational space can predict English speaking participants results for the studied task.We find that several models are predictive of humans. Surprisingly, a multilingual model—which is not trained to recognize English phonemes—and a short-duration acoustic event model—which is not trained to recognize phonemes at all—are far more predictive than a standard speech recognizer. We argue that the speech recognizer overfits on the language it has been trained on—English—and organizes its representational space in a way that is different from English-speaking listeners' perceptual space.

Perceptimatic English Benchmark

Experimental Task We assess the perception of short phone sequences. We use an ABX discrimination task. Human participants hear three stimuli and are asked to identify which one of the first two stimuli (A or B) is more similar to the third (X). The experimenter always identifies a correct answer—in this case, by making A and B instances of two different phonetic categories, and X another example of one of the two. The accuracy of listeners' responses to a given *triplet* (combination of specific stimuli into an A–B–X item) gives a measure of the discriminability of the categories to which A and B belong.

Stimuli We construct triplets in which A, B, and X are each sequences of three consecutive phones (triphones) extracted from running speech, where the phonemic-level transcription indicates that only the centre phone differs between A and B (for example, [seɪk]–[soʊk]). Both English and French stimuli are extracted from the subset of the LibriVox audio book collection used as evaluation stimuli in the Zero Resource Speech Challenge (see **Related work** below). We control the context in order to avoid mismatching different contextual allophones. We incorporate this context into the stimuli in order to avoid making the stimuli too short. The stimuli are not an arbitrary subset, but are a nearly-balanced subset selected by hand, taking the phonemic retranscription of the corpus as a starting point, and performing a manual verification to select clear examples (see below).

We exclude phones (or phones in certain neutralizing contexts) which we expected might be subject to a merger for some listeners, or which were sufficiently marginal that the corpus transcriptions were unlikely to be reliable. Not all phone comparisons occur, nor do all phone comparisons occur in the same contexts, or with the same set of speakers: we (native English and French listeners) selected the stimuli by hand out of the very large set of constructible triplets to maximize the phonetic similarity of the probe's centre phone to that of the correct answer, and to minimize phonetic differences in the surrounding contexts. This is critical when extracting stimuli from natural speech: transcriptions are not always accurate, and a three-phone window is not sufficient to guarantee which of the many possible contextual variants each transcribed phone really corresponds to.[2]

For each ABX triplet, the reference stimuli, A and B, are uttered by the same speaker, in order to avoid listeners' responding on the basis of speaker differences, while the probe, X, is uttered by a different speaker, to encourage listeners to focus less on minor acoustic details. The delay between A and B is 500 milliseconds, and between B and X 650 milliseconds, as pilot subjects reported having difficulty recalling the reference stimuli when the delays were exactly equal.

In total, the stimuli consist of 5202 triplets (2214 from English), making 461 distinct centre phone contrasts (212 English, 249 French), in a total of 201 distinct contexts (118 English, 83 French), with most phone comparisons appearing in three contexts each (a total of 47 English contrasts appear in either one, two, or four contexts). The speakers used (15 English, 18 French) have, in our assessment, standard, broadcast-type American English/Metropolitan French pronunciations. Each set of three stimuli (triplet) appears in four distinct items, corresponding to orders AB–A (that is, X is another instance of the three-phone sequence A), BA–B, AB–B, and BA–A.

We note a few things about the construction of this test. First, while speaker variability was introduced in order to prevent listeners from attending to acoustic details, the delay between stimuli is still relatively short, meaning that listeners need not rely heavily on memory, and will thus still have reasonable access to detail. The stimuli are also short, and are sometimes not cut at syllable boundaries, so that listeners may not treat them as fully speech-like. The fact that the A and B stimuli are from the same speaker may also attune listeners to small differences between those two stimuli, potentially thus attuning them to low-level differences overall. By testing individuals on stimuli of this kind, we expect that we will obtain a profile of low-level phonetic/auditory discrimination.

In order to understand the gap between human listening and typical speech recognition models—and, more generally, any statistical model of acoustics used in automatic speech processing, serving either applied or fundamental science purposes—we will require a broad spectrum of different kinds of tests, in different listening modes. The stimuli we use here, which are extracted from running speech, are designed with the express purpose of putting current speech processing models in ecological listening conditions, and thus represent a narrow starting point for the broader testing architecture. As a speech perception experiment, the results are difficult to interpret, for several reasons. While the first and last phones were, in principle, held constant across each stimulus triplet, in reality, it is very difficult to get phonetically well-matched contexts in natural speech. Although the stimuli were selected by hand to minimize the differences due

[2]The full set of English centre phones included in at least one item is [æ ɑ b d ð eɪ ɛ f g h i ɪ k l m n ŋ oʊ p ɹ s ʃ t tʃ u ʊ v ʌ w z]. The full set of French phones included is [a ɑ̃ b d e ɛ ɛ̃ f g i j k l m n ɲ o ø ɔ ɔ̃ p ʁ s ʃ t u v w y z ʒ]. For the full list of pairs and contexts tested, see the online repository.

to surrounding context, they are far from perfectly controlled, which means that the target (centre) phone is not the only thing that will drive human listeners' decisions. Thus, grouping the stimuli by centre phone contrast for analysis is risky. We make our comparisons with models only on an individual triplet stimulus level, so as not to suppose that the only source of difference is the centre phone. We assess the global predictiveness of the relative distance or similarity produced by each model for the pairs AX versus BX, for the probability of an accurate response in the human listeners. The overall predictiveness gives an indication of the similarity of models' representational spaces to the perceptual encoding used by human listeners in a low-level speech listening task.

Reference Data Collection The data set includes 91 participants reporting English as the language to which they were primarily exposed before the age of eight. They performed the task on Amazon Mechanical Turk (US participants) with the LMEDS software (Mahrt, 2016) and were paid for participation.[3] We asked participants to use headphones, to do the task in a quiet environment, and to check the sound volume before the experiment began. 15 additional participants were tested but did not meet the language background requirements, and 65 were rejected for failing at least three out of twelve catch trials or not finishing the task.[4]

For testing, items were counterbalanced into lists of 190 triplets per participant, such that no participant was tested twice on the same contrast, and such that the combination of speakers was not predictive of the right answer. Each stimulus was tested three times, so that most contrasts were tested at least 36 times. Participants responded as to which of the two reference stimuli the probe corresponded to on a six-point scale, ranging from *first for sure* to *second for sure,* with two intermediate degrees of certainty in favour of each reference stimulus. The benchmark includes both these responses and a binarized version, taking into account the participant's choice but not their reported certainty. Here we report only analysis of the binarized responses to avoid questions about how to model participants' use of the scale (preliminary analyses on the scaled responses indicate that the results are qualitatively the same).

The results we obtained are consistent with the assessment that listeners tap into a low-level phonetic/auditory mode of listening.While many of the difficult phone pairs seem reasonable, others do not: among the more difficult English contrasts for listeners are English [f]–[v], which should not be particularly difficult, and French [f]–[y], which should be trivially easy. The reason must be that the relevant set of items did not highlight the centre phone contrast. Furthermore, the fact that the locus of contrast was not always apparent might also have led listeners to attend to acoustic detail across the whole experiment. Again, we use this kind of behaviour to justify a stimulus triplet-level analysis.

In spite of the apparent fact that listeners made use of low-level phonetic detail, the results are not driven by purely auditory mechanisms completely irrelevant to speech. We argue this on the basis of the fact that there is a native language effect, such that listeners are globally better at native-language (English) stimuli than non-native (French) stimuli. This suggests that linguistically relevant processing is still revealed by this test.

Generating Model Predictions

For each experimental stimulus, we apply a model to the audio file and extract that model's representation of the stimulus (see below for examples). We use these representations to compute distances $d(\text{Target},\text{X})$, between the probe and the correct matching stimulus, and $d(\text{Other},\text{X})$, between the probe and the other reference stimulus, to generate a degree of correct discriminability $\delta = d(\text{Other},\text{X}) - d(\text{Target},\text{X})$. If $\delta > 0$, then the model treats the probe as being more similar to the correct than to the incorrect answer. Our goal is to assess whether humans' perceived similarity matches the model's distances (δ values). Humans' responses are stochastic, and need not use a threshold at the point of maximal perceived similarity. This leads us to use a binomial generalized linear model with an intercept parameter (see Section). This is similar to generating confusion matrices at the level of individual stimuli, and comparing those generated by acoustic model distance with those given by human accuracies.

Using δ values is not the only possible linking hypothesis, but it is broadly applicable, and allows for a distance function to be selected that is appropriate to the type of representation being tested. All the models we consider in this paper yield representations of variable length (they output vector sequences—one vector per time frame—and the stimuli are not all of the same duration). Thus, we use distance functions based on dynamic time warping. Dynamic time warping takes two sequences C and D as input, as well as a function γ for comparing pairs of sequence elements. It aligns C and D by matching the elements of one to the other so as to minimize the sum of $\gamma(c,d)$ for all matched elements (c,d). Each element of C must to be matched with at least one element of D, and alignments must respect temporal order. Here we calculate distances between stimuli $C = c_1, c_2, ...c_p$ and $D = d_1, d_2, ...d_q$ as:

$$d(C,D) = \frac{\sum_{c_i,d_j \text{ are matched}} \gamma(c_i,d_j)}{max(p,q)} \quad (1)$$

For the models analyzed here, we take γ to be either a symmetrised Kullblack–Leibler divergence[5] (for models that output probability vectors), or a cosine distance. Where $\mathbf{x}$ and $\mathbf{y}$

[3] Kleinschmidt & Jaeger, 2015 made a detailed comparison of data from an in-lab speech perception experiment with a Mechanical Turk replication and found a close correspondence between the results.

[4] The catch trials consisted of additional, highly distinct three-phone ABX stimuli, including several which required participants to distinguish *cat* from *dog*.

[5] We replace zero elements with a very small constant to avoid division by zero.

are N-dimensional vectors, they are defined as:

$$\gamma_{KL}(\mathbf{x},\mathbf{y}) = \frac{1}{2}\left[\sum_{i=1}^{N} x_i \log(\frac{x_i}{y_i}) + \sum_{i=1}^{N} y_i \log(\frac{y_i}{x_i})\right] \quad (2)$$

$$\gamma_{cos}(\mathbf{x},\mathbf{y}) = \frac{1}{\pi}\arccos\left(\frac{\sum_{i=1}^{N} x_i y_i}{\sqrt{\sum_{i=1}^{N} x_i^2}\sqrt{\sum_{i=1}^{N} y_i^2}}\right) \quad (3)$$

Experiments

We apply the methods described in the previous section. Unless stated, we take γ to be the cosine distance (3).

Dirichlet Process Gaussian Mixture Model We evaluate a Dirichlet process Gaussian mixture model (DPGMM) as proposed by Chen, Leung, Xie, Ma, and Li (2015). Given a training set of speech recordings in a language, the model performs non-parametric Bayesian clustering on the entire database, treated as an unordered collection of instantaneous acoustic feature vectors (see **Mel Filterbank Cepstral Coefficients** below). It models short-duration acoustic events. A fitted model consists of an optimal set of Gaussian distributions—typically several hundred. The model thus preserves fine-grained temporal and acoustic detail, while still modelling a specific language. It does not use phoneme labels. Passing over a new sample at a fixed analysis rate (in our case, analyzing 25 milliseconds of signal every 10 milliseconds), each instant of signal is mapped to a vector of posterior probabilities over the Gaussians in the model. We take γ to be the symmetrized KL divergence (2). We apply the English model described in Millet, Jurov, and Dunbar (2019), trained on 34 hours of English speech taken from the LibriVox dataset (no overlap with the stimuli or speakers in Perceptimatic).

Bottleneck features We evaluate three models proposed in Silnova et al. (2018). These *bottleneck* models are trained to label speech with *phone states*. Phone states are temporal analysis units used by certain speech recognizers: each phone of the language is modelled as having (in the typical three-state model) a beginning, middle, and end state, each with different acoustic properties. The bottleneck models are trained on speech data labeled annotated with an attribution to phone states. They are neural networks trained to predict the phone state associated with a given instant of speech, on the basis of its acoustic features, accompanied by 310 ms of surrounding context. This model is thus optimized to predict a more temporally fine-grained version of standard phoneme labels. "Bottleneck" refers to a hidden layer that has significantly lower dimension than the other layers. The features we use are the contents of this layer, for each instant of signal. We evaluate *English monophone, English triphone,* and *multilingual* models.[6] The English monophone model is optimized to predict states for English phonemes. The English triphone model is optimized to predict states for contextual allophones. The multilingual model is trained on data from seventeen phonetically diverse languages (not including English), optimized to label phoneme states in any of these languages (if the same sound belongs to different inventories, it is treated as distinct, for a total of 1032 possible phonemes).

DeepSpeech DeepSpeech (Hannun et al., 2014) is a neural automatic speech recognition model used in the Mozilla speech tools.The model uses bi-directional recurrent units, which integrate information both forwards and backwards in time, to predict text transcriptions (sequences of letters, not phones) from speech. We can examine the state of any of its several internal layers corresponding to any instant of signal. After scoring each layer on its performance on the phone discrimination metric described in Dunbar et al. (2017) (on which the PEB is based), we found that layer five was optimal. We thus analyze the outputs from that layer. The model has a training objective related to that of the English bottleneck models (predicting text), but the recurrent units allow it to model long distance temporal dependencies, and the units to be predicted are graphemes, which are more similar in their temporal granularity to phonemes than to phone/phoneme-states. We use Mozilla DeepSpeech 0.4.1 [7], which is trained on the Fisher (Cieri, Miller, & Walker, 2004) and Switchboard (Godfrey, Holliman, & McDaniel, 1992) telephone corpora and the LibriSpeech audio book corpus (Panayotov, Chen, Povey, & Khudanpur, 2015). The model achieves an 8.26% word error rate on the LibriSpeech clean test evaluation.

Articulatory Reconstruction To explore whether similarities at the level of articulation are more predictive of humans' behaviour, we evaluate a neural *articulatory reconstruction* model (Parrot, Millet, & Dunbar, 2019), trained to reconstruct continuous electromagnetic articulography (EMA) coil position trajectories from speech recordings (tongue body, tongue tip, tongue dorsum, upper lip, lower lip, lower incisor). The model is trained on the EMA-IEEE corpus (Tiede et al., 2017), approximately six hours of read English speech, paired with EMA recordings, from eight speakers.

Mel Filterbank Cepstral Coefficients We use Kaldi (Povey et al., 2011) to extract 13 Mel filterbank cepstral coefficients (MFCC): one vector every 10 milliseconds, each analyzing 25 milliseconds of signal. These audio representations, used standardly as input to speech recognition, are the result of a low-resolution spectral analysis and a discrete cosine transformation. We add the first and second derivatives, for a total of 39 dimensions, and apply mean-variance normal-

[6] Referred to by Silnova et al. (2018) as *FisherMono, FisherTri,* and *BabelMulti*.

[7] https://github.com/mozilla/DeepSpeech/releases/tag/v0.4.1

	PEB	GMM	DS	BEnM	BEnT	BMu	Art	MFCC
En	79.5	88.3	89.5	91.2	90.3	88.9	77.3	78.6
Fr	76.7	82.0	80.2	87.6	88.8	88.5	70.1	78.3

Table 1: Percent accuracies for humans (PEB) and models (the bigger the better). GMM is for DPGMM, DS for DeepSpeech. BEnM, BEnT and BMu are (in order) for monophone English, triphone English and multilingual bottleneck models. Art is for articulatory reconstruction.

ization over a moving three-second window. This approach, like the multilingual bottleneck features, does not specifically model English; unlike that model, this is a fixed transformation, not tuned to any language, or indeed to speech at all.

Results

Performance on the Experimental Task We compute the mean accuracies[8] for each of the models, scoring stimuli as correct where $\delta > 0$. The results in Table 1 indicate that the models' performance is generally better than the human listeners in the PEB. This implies that, to the extent that any of these models accurately captures listeners' perceived discriminability, listeners' behaviour on the task, unsurprisingly, cannot correspond to a hard decision at the optimal decision threshold. The results also indicate a small native language effect—a decrease in listeners' discrimination accuracy for the non-English stimuli. Such an effect is also captured by all the models trained on English. We observe that some models show native language effects numerically much larger than human listeners, a point we return to below.

Prediction In order to see which model best predicts the human results,[9] we fit probit regression models with a coefficient for the δ discriminability score corresponding to the given model. The dependent variable is whether the trial response was correct (1: accurate, 0: inaccurate). To correct for effects that are not of interest, the models each also include a coefficient for whether the correct answer was A or B, a coefficient for the position of the trial in the experimental list, and a coefficient for participant.

We use differences in log likelihood for model comparison, obtaining confidence intervals by repeatedly drawing balanced subsamples ($N = 43358$): for each stimulus, we draw three observations without replacement. The results, in Table 2, show that the most predictive approaches are short-term acoustic event modelling (DPGMM) and bottleneck phone state predictors, with the English monophone (phoneme) predictor model showing non-significantly poorer performance

[8]Scoring accuracy first by stimulus, then averaging by contrast, then overall.

[9]Here we report results combining the English (native) and French (non-native) stimuli. In the interest of analyzing stimuli that are maximally ecological for the models tested, we also analyzed the results of the native-language perception task only. The model comparisons are qualitatively identical, so we omit the results in the interest of space.

	BMu	BEngT	BEngM	MFCC	DS	Art
GMM	3	9	28	**204**	**249**	**257**
BMu		6	24	**202**	**246**	**254**
BEngT			19	**196**	**241**	**248**
BEngM				**177**	**222**	**229**
MFCC					**45**	**52**
DS						8

Table 2: Mean of resampled differences in log likelihood between models. Models are ordered by column and by row in descending order of their performance, with better models on the left/top. Positive numbers indicate that the model indicated in the given row is better than the model indicated in the column. Bolded results have 95% confidence intervals that exclude zero. GMM is for DPGMM, DS for DeepSpeech. BEnM, BEnT and BMu are (in order) for monophone English, triphone English and multilingual bottleneck models. Art is for articulatory reconstruction.

than the allophonic and multilingual ones. This means that these models have representational spaces that are the closest to human perceptual space of all the tested models.

Discussion

Behaviour on our speech discrimination benchmark are best predicted by the DPGMM's short-duration acoustic event modelling and the three bottleneck phone state classification models, consistent with Millet et al. (2019) and Jurov (2019). These do substantially better than generic audio features. These means that their representational space is the closest to the perceptual space used by humans in this task. Two of the bottleneck models are trained to predict English phoneme/allophone states, but the multilingual model is not trained on English, which makes its performance all the more surprising. The DPGMM model, which, although trained on English, models 25 millisecond acoustic events into combinations of hundreds of detailed acoustic categories of its own devising, and is thus much more temporally and acoustically fine-grained than typical phonetic annotation.

The articulatory reconstruction model is not very predictive of human behaviour. The likely reasons are simple. First, predicting articulatory parameters for novel speakers is difficult, and this model is far from having state-of-the-art performance. Second, the model does not predict a complete set of articulators. It is thus unsurprising that, when scored on the experimental task, this model is worse than even the acoustic features.

The continuous speech recognizer (DeepSpeech) is also bad at predicting human behaviour, but, unlike the articulatory reconstruction, performs well on the experimental task. This model is different from the English bottleneck models in three ways. First, it is in principle capable of taking into account longer temporal dependencies than the finite 310 ms window used by the bottleneck model. Second, it is optimized not to predict phonemes or allophones, but or-

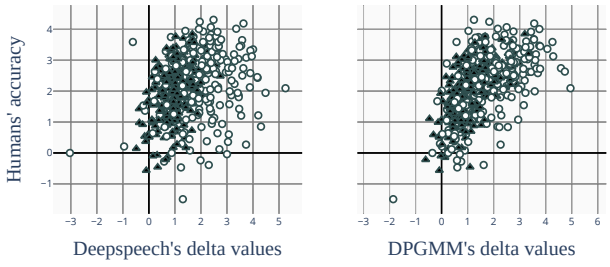

Figure 1: Normalized human accuracy versus normalized δ for (left) the DeepSpeech speech recognizer and (right) the DPGMM short-duration acoustic model. White circles are English stimuli and black triangles are French stimuli.

thographic (letter) transcriptions. These are not equivalent, since English orthography is not completely transparent: distinct sequences of phones correspond to distinct sequences of letters (thus, allow for a high score on the experimental task), but the representation's distances may capture similarities and differences exclusively found in spelling. Finally, the bottleneck models are optimized to predict temporally more fine-grained sequences (distinguishing between beginning, middle, and end states for each phoneme or allophone), while transitions between English letters roughly correspond to transitions at a level of temporal granularity similar to that of the segment.

DeepSpeech shows the largest discrepancy between the English and French stimuli (larger than English listeners'). This difference is clear from Figure 1 (left), which plots DeepSpeech's δ discriminability scores against listeners' averaged accuracy,[10] colour-coded for whether the items are native (English) or non-native (French). We observe a clear separation in the distributions of DeepSpeech's predicted discriminability for the English stimuli (concentrated on the right-hand part of the graph, where the model is better) versus French stimuli. There is also a (statistically significant) separation between native and non-native stimuli for humans, but it is small enough as to be visually far less salient. Correspondingly, the DPGMM model (right) also shows a smaller separation between the two types of stimuli (though still larger than English listeners': see Table 1). Furthermore, DeepSpeech's δ values for the English-language stimuli are also more homogeneous than the DPGMM's, with a less pronounced slope from difficult to easy stimuli. It would seem that optimizing on the task of predicting English grapheme sequences leads DeepSpeech to attend to, or ignore, very different acoustic information than human listeners, at least in the context of this low-level speech discrimination task.

[10]Here we group by centre phone pair in order to have more resolution in our measure. An equivalent analysis making use of subjects' gradient responses is qualitatively similar.

Related work

Our data set is in the spirit of other cognitive benchmarks for artificial intelligence (syntax: Warstadt et al., 2019; intuitive physics: Riochet et al., 2018; question answering: Kwiatkowski et al., 2019). In speech perception, the idea of matching human behaviour is not new (Kleinschmidt & Jaeger, 2015; Feldman & Griffiths, 2007; Schatz, Feldman, Goldwater, Cao, & Dupoux, To appear; Schatz, Bach, & Dupoux, 2017; Schatz & Feldman, 2018), and is an echo of the literature on modelling phonetic learning, most notably Guenther and Gjaja (1996), who qualitatively compared their modelled distances to similarities reported in the literature for human listeners. To our knowledge, the only previous work providing stimuli, human responses, and recommendations for generating predictions at the individual stimulus level with a wide range of models is Millet et al. (2019). Those stimuli only tested a narrow range of cross-linguistic phone contrasts, however, and were non-words read in a word-list style, rather than extracts of natural, running speech.

The PEB stimuli are drawn from the evaluation for the Zero Resource Speech Challenge 2017 (Dunbar et al., 2017), widely used in evaluating unsupervised speech models. The PEB complements this existing measure (the existing measure is not scored against humans), and can be applied to any model already tested on the Zero Resource Speech Challenge 2017 evaluation.

Summary of Contributions

We have presented the **Perceptimatic English Benchmark**, an open benchmark for computational models of human speech perception made up of English and French stimuli that are ecological for typical speech models. It is the only open data set we know of that systematically probes a wide range of phone contrasts and that enable us to easily compare English speaking humans with computational models for a low-level speech discrimination task. We have shown, for the first time, that a standard speech recognizer trained on English recordings is not predictive of English speaking human phone classification behaviour, while models not optimized to recognize English phonemes are (a quasi-universal phone classifier and a model of short-duration acoustic events). The multilingual model is easy to use off-the-shelf,[11] and we recommend it to researchers needing an estimate of perceptual distance.

Acknowledgements

This research was supported by the École Doctorale Frontières du Vivant (FdV) – Programme Bettencourt, and by 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.

[11]https://coml.lscp.ens.fr/docs/shennong/.

References

Chen, H., Leung, C.-C., Xie, L., Ma, B., & Li, H. (2015). Parallel inference of Dirichlet process Gaussian mixture models for unsupervised acoustic modeling: A feasibility study. In *INTERSPEECH-16*.

Cieri, C., Miller, D., & Walker, K. (2004). The fisher corpus: a resource for the next generations of speech-to-text. In *Lrec*.

Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., ... Dupoux, E. (2017). The Zero Resource Speech Challenge 2017. In *2017 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)* (pp. 323–330).

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Elman, J., & McClelland, J. (2015). Exploiting the lawful variability in the speech wave. In J. Perkell & D. Klatt (Eds.), (Vol. 335, pp. 71–90). Hillsdale, NJ: Erlbaum.

Feldman, N. H., & Griffiths, T. L. (2007). A rational account of the perceptual magnet effect. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 29).

Godfrey, J. J., Holliman, E. C., & McDaniel, J. (1992). Switchboard: Telephone speech corpus for research and development. In *[proceedings] icassp-92: 1992 ieee international conference on acoustics, speech, and signal processing* (Vol. 1, pp. 517–520).

Guenther, F., & Gjaja, M. (1996). The perceptual magnet effect as an emergent property of neural map formation. *The Journal of the Acoustical Society of America*, *100*(2), 1111–1121.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., ... others (2014). Deep speech: Scaling up end-to-end speech recognition. *arXiv preprint arXiv:1412.5567*.

Jurov, N. (2019). *Phonetics or Phonology? Modelling Non-Native Perception*. Unpublished master's thesis, Université Paris Diderot, Paris, France.

Kleinschmidt, D., & Jaeger, T. F. (2015). Robust speech perception: Recognize the familiar, generalize to the similar, and adapt to the novel. *Psychological Review*, *122*(2), 148–203.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., ... others (2019). Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, *7*, 453–466.

Mahrt, T. (2016). *LMEDS: Language markup and experimental design software.*

McClelland, J., & Elman, J. (1986). Interactive processes in speech perception: The TRACE model. *Cognitive Psychology*, *18*, 1–86.

Millet, J., Jurov, N., & Dunbar, E. (2019). Comparing unsupervised speech learning directly to human performance in speech perception..

Norris, D., & McQueen, J. (2008). Shortlist B: a Bayesian model of continuous speech recognition. *Psychological Review*, *115*(2), 357–395.

Panayotov, V., Chen, G., Povey, D., & 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)* (pp. 5206–5210).

Parrot, M., Millet, J., & Dunbar, E. (2019). Independent and automatic evaluation of acoustic-to-articulatory inversion models. *arXiv*, arXiv–1911.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., ... others (2011). The Kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU)*.

Riochet, R., Castro, M. Y., Bernard, M., Lerer, A., Fergus, R., Izard, V., & Dupoux, E. (2018). Intphys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*.

Scharenborg, O., Norris, D., ten Bosch, L., & McQueen, J. (2005). How should a speech recognizer work? *Cognitive Science*, *29*, 867–918.

Schatz, T., Bach, F., & Dupoux, E. (2017). ASR systems as models of phonetic category perception in adults. In *Proceedings of the 39th Annual CogSci Meeting*.

Schatz, T., & Feldman, N. (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* (pp. 1–4).

Schatz, T., Feldman, N., Goldwater, S., Cao, X. N., & Dupoux, E. (To appear). Early phonetic learning without phonetic categories: Insights from machine learning. *Proceedings of the National Academy of Sciences*.

Silnova, A., Matejka, P., Glembek, O., Plchot, O., Novotnỳ, O., Grézl, F., ... Cernockỳ, J. (2018). But/phonexia bottleneck feature extractor. In *Odyssey 2018: The speaker and language recognition workshop* (pp. 283–287).

Tiede, M., Espy-Wilson, C. Y., Goldenberg, D., Mitra, V., Nam, H., & Sivaraman, G. (2017). Quantifying kinematic aspects of reduction in a contrasting rate production task. *The Journal of the Acoustical Society of America*, *141*(5), 3580-3580. Retrieved from `https://doi.org/10.1121/1.4987629` doi: 10.1121/1.4987629

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. (2019). Blimp: A benchmark of linguistic minimal pairs for english. *arXiv preprint arXiv:1912.00582*.