Comparing unsupervised speech learning directly to human performance in speech perception

Juliette Millet (juliette.millet@cri-paris.org)

Nika Jurov (nika.jurov@gmail.com)

Ewan Dunbar (ewan.dunbar@univ-paris-diderot.fr)

Laboratoire de Linguistique Formelle (CNRS – Université Paris Diderot – Sorbonne Paris Cité)
and Cognitive Machine Learning (ENS – CNRS – EHESS – INRIA – PSL Research University)
Paris, France

Abstract

We compare the performance of humans (English and French listeners) versus an unsupervised speech model in a perception experiment (ABX discrimination task). Although the ABX task has been used for acoustic model evaluation in previous research, the results have not, until now, been compared directly with human behaviour in an experiment. We show that a standard, well-performing model (DPGMM) has better accuracy at predicting human responses than the acoustic baseline. The model also shows a native language effect, better resembling native listeners of the language on which it was trained. However, the native language effect shown by the models is different than the one shown by the human listeners, and, notably, the models do not show the same overall patterns of vowel confusions.

Keywords: linguistics; language acquisition; machine learning; speech recognition

Introduction

Comparing cognitive models with human behaviour often involves some idealization. The ideal comparison between a model and a human behavioural experiment would simply have the model "participate" in the experiment, exposed to the same stimulus files as are presented to the humans, responding as if it were just another human subject. Its responses would be compared to human subjects' on a stimulus-by-stimulus level. This ideal is reached only rarely (for example, Riochet et al., 2018). Most settings either simplify the stimuli given to models (for example, showing images of objects to human participants, but providing the model instead with a discrete input indicating whether the object was a dog or a cat, as in Xu & Tenenbaum, 2007), or compare highly aggregated results rather than predictions on individual stimuli (for example, Gulordava et al., 2018). These simplifications, while often essential, may mask aspects of the real task which have a major impact on the results.

Meanwhile, a large body of recent research has proposed to evaluate acoustic models trained on speech databases, particularly those trained in an unsupervised way, using an *ABX phone discrimination task* (Schatz et al., 2013). This evaluation considers pairs of speech stimulus items (A and B) coming from two different phonemic categories, assessing whether the model's representation of a third stimulus (X) is more similar to its representation of A or of B.

While this task is analogous to the standard human ABX perception task, a direct comparison of the two to evaluate models or better understand human behaviour has not yet been done. We propose a direct, stimulus-by-stimulus comparison of an acoustic model with human perception in an ABX perception task. Additionally, the stimuli for our task come from two different languages. We examine the behaviour of human subjects, and trained models, for whom one of the languages is a second language (L2). Previously, unsupervised acoustic models have typically been evaluated by assessing how well they discriminate phonemes of the language on which they are trained (L1), their objective being to reach perfect discrimination of all pairs of phonemes in the L1 (Schatz et al., 2013; Versteegh et al., 2015). A few studies have investigated patterns of L2 discrimination in acoustic models, looking at overall accuracy on phonemic contrasts from languages other than the training language. But their conclusions have been based on qualitative summaries of the behaviour of the models, with no human reference data on the same stimuli (Schatz et al., 2017; Schatz & Feldman, 2018).

A stimulus-by-stimulus comparison of an acoustic model with human performance on a speech perception task might reveal major differences between the two. If a trained acoustic model is seen as an acoustic baseline, the comparison will highlight aspects of human speech perception which are surprising given properties of the signal alone. On the other hand, if the goal of the acoustic model is to be human-like, such a comparison shows us where the model falls short.

We train an unsupervised acoustic model which is known to perform globally well on corpus-based ABX evaluations (Chen et al., 2015). We train the model on English and French corpora. We expose both the English-trained model and the French-trained model to novel, experimental stimuli. We evaluate the models' ABX discrimination accuracy. We give English and French human native listeners the same task.

Our results show that the model is globally more predictive of the human results than a baseline based on low-level acoustic features. The model also shows a native language effect: when trained on French, its error pattern is more like French native speakers', and similarly for English. However, we analyze these error patterns, and show that the native language effects shown by the models, while globally predictive, differ importantly from those shown by the human participants.[1]

[1] All modelling code, analysis code, stimuli, and anonymized

Methodology: Human ABX evaluation

In an ABX paradigm, participants hear three sounds in sequence, and indicate which of the first two sounds (A or B) is more similar to the last (X), a sound always drawn from the same category (for example, phoneme) as either A or B. The task is intended to tap the perceptual similarity between A and X, on the one hand, and B and X, on the other, to assess the overall distinctness of the categories A and B belong to.

We develop stimuli to test cross-linguistic (English/French) perception of vowels in an ABX discrimination paradigm. Within each stimulus triplet, A and B always consist of CVC non-words contrasting one English vowel with one French vowel, with the flanking consonants held constant. We use the American English vowels [ɪ], [ʌ], [ʊ], and [æ], and the Hexagonal French vowels [a], [ɔ], [ɛ], [i], [u], [y], and [œ].[2] Only consonants appearing in both languages are used: [v], [z], [s], [ʃ], [f], in both consonant positions, and, additionally, [p], [b], [g], and [k] in coda.[3] While the stimuli are designed to differ only in the vowel, there are inevitable phonetic differences in the realization of these consonants across the two languages, which may provide additional cues to the correct answer. Real words in either language are excluded. For details of stimulus construction, see **Experiments: Humans** below.

We expect that human listeners will vary in their discrimination ability, with triplets like [vip]–[væp]–[vip] being generally more difficult than more acoustically similar triplets such as [vʌp]–[vɔp]–[vʌp]. We also expect cross-linguistic differences, with English listeners doing better than French listeners on acoustically similar contrasts which do not exist in French, such as [i]–[ɪ]. We examine the patterns of confusions shown by both listener groups, and present the same experimental stimuli to models trained on English and on French, to evaluate the models' internal representations.

Methodology: Model ABX evaluation

Unsupervised acoustic models are models that learn representations of speech by exposure to speech without associated phonemic category labels. They can be seen as learning the organization of a perceptual space for speech.

We train a Dirichlet Process Gaussian Mixture Model (DPGMM) as an acoustic model. It is a non-parametric Bayesian clustering model. It finds, in an unsupervised way, a set of multi-dimensional Gaussian distributions appropriate to cluster the observations (here acoustic features). It adapts its number of Gaussian distributions automatically depending on the training data. The computations needed by the model training can be parallelized (Chang & Fisher III, 2013), making training on a reasonable amount of speech data possible. The resulting trained model (learned set of Gaussian distributions) can then be applied to any new speech example, yielding a sequence of probability vectors that can be seen as the model's perceptual representation of the example. In this way, the model can be seen as learning the organization of a perceptual space. Chen et al. (2015) applied parallel DPGMM training and achieved the best performance in the 2015 ZeroSpeech Challenge, a machine learning challenge seeking state-of-the-art unsupervised acoustic models (Versteegh, Anguera, Jansen, & Dupoux, 2016).

The representations we extract from the DPGMM model are posteriorgrams. A speech signal consists of a sequence of audio frames: for a sequence of k audio frames, a posteriorgram is a sequence of k vectors. The vector $\mathbf{x_i} = (x_1, x_2, ...x_N)$ gives the probabilities of the i^{th} frame having been generated by each of the model's N learned Gaussian distributions.

Performing ABX evaluation of an encoding learned by an acoustic model relies on extracting the representations of triplets of stimuli (A, B, and X), and computing the distance $d(\mathrm{A},\mathrm{X})$, between A and X, and $d(\mathrm{B},\mathrm{X})$, between B and X. X is of the same category as either A or B. Taking A to be the correct answer, we compute $\delta = d(\mathrm{B},\mathrm{X}) - d(\mathrm{A},\mathrm{X})$. If $\delta > 0$, we can consider the model to have chosen A; if $\delta < 0$, we consider it to have chosen B. In previous work evaluating acoustic models with this method (Versteegh et al., 2015; Dunbar et al., 2017), the percentage of correct responses for each pair of categories is tabulated, and these averages are combined into a global ABX discriminability score.

Because it relies only on computing distances, the model ABX evaluation is applicable to a broad variety of learned representations. It can be applied to posteriorgrams, but also to Mel-frequency cepstral coefficients (MFCCs), a compact representation of acoustic cues derived from the spectrum, commonly used to train ASR models. We train our models here on MFCC inputs, and MFCCs also serve as our low-level acoustic baseline (each audio frame is a MFCC vector).

The distance function most appropriate for the comparison may vary as a function of the type of representation. Because the representations we evaluate contain one vector per audio frame, differing-length stimuli will have different-length representations. To deal with those differences, we follow previous literature in the domain and use dynamic time warping (DTW) to align the sequences (see Senin, 2008 for a review). This algorithm computes an optimal match between two sequences based on a secondary distance function used for comparing individual elements across the two sequences (individual vectors in the speech representations). Every frame in each of the two representations is matched with at least one frame in the other representation, following the order of each sequence. The final distance between the two sequences is the mean of the distance between the matched frames.

As secondary distance functions, we use the same frame-level distances as in previous evaluations of DPGMM acous-

data for this paper are available in the following online repository: `https://github.com/geomphon/CogSci-2019-Unsupervised-speech-and-human-perception.git`.

[2]This reduced set of vowels is constructed with special attention to French native listeners' perception of the English vowel [ʌ]. Previous research shows (Peperkamp, 2015) that French native listeners identify this vowel with a number of different French vowels, suggesting that a fair number of pairs will be difficult for subjects.

[3]Stops are excluded in onset position because of the marked differences between English and French VOT.

tic models. For MFCC representations, we use the cosine distance. For N-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, it is defined as:

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

For comparing the posteriorgrams of our trained models, we use the symmetrized Kullback–Leibler (KL) divergence. For positive[4] N-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$, the symmetrized KL-divergence between x and y is:

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

Although this model ABX task is inspired by a speech perception task, the test is different from a typical speech discrimination experiment in an important way. By tabulating the proportion of triplets with $\delta > 0$ (correct), gradient information about individual stimuli triplets is lost. Such a test cannot measure how well separated or "discriminable" individual speech stimuli are, but only the separation of a pair of categories A and B. Rather than directly using model ABX discriminability scores, we relate human discrimination of individual stimuli to δ; see below, and see also Schatz, 2016.

Experiments: Humans

The stimuli were recorded in a carrier phrase. Six speakers read the stimuli in an anechoic chamber. Two were early bilinguals of American English and Hexagonal French, both female, and read both the English and the French vowel stimuli. Both had extensive exposure to both languages throughout most of their early and adult lives, and regularly used both languages. These stimuli were used for A and B. The other four speakers were male: two North American English natives, who read the English stimuli, and two Hexagonal French natives, who read the French stimuli. Their productions were used as X. Phonetically trained listeners (one French and one English native), listened to the stimuli in isolation and verified that they were native-like in the target language and corresponded to the intended vowel.

All A and B pairs were cross-language comparisons. If A was a French stimulus, B was English, and vice versa. The A and B speakers always differed. The experiment used 500 ms silence for both the A–B and B–X intervals.

The final set of stimuli consisted of 112 triplets, matched to the same intensity, downsampled to 16000 Hz. The list was a subset of the complete set of possible triplets, optimized to balance combinations of speaker, vowel pair, consonantal context, and whether A or B was the correct answer. Each vowel pair appeared four times, factorially combining which of the two vowels was the correct answer, and whether the correct answer was presented first (A) or second (B).

The task was performed on Amazon Mechanical Turk with the LMEDS software (Mahrt, 2016), with participants from the United States and France. Listeners were paid for participation. Previous research shows that Mechanical Turk can successfully be used for speech perception tasks, and that results are comparable to a lab setup (for example, Kleinschmidt & Jaeger, 2015). We asked the participants to use headphones, to do the task in a quiet environment, and to check the sound volume before the experiment began.

A total of 144 participants were tested, 72 in France and 72 in the United States. We filter out those who did not finish the task, did not report English or French as their first language, had previously taken a linguistics class, failed two out of three catch trials[5] or reported hearing or vision problems. In the end, there were 63 English and 55 French participants.[6]

Experiments: Models

To build the models for comparison with the human experiment, we train the DPGMM on the same LibriVox audio book source corpora used to construct the English and French data sets in the 2017 ZeroSpeech Challenge (Dunbar et al., 2017). We use a different subset of the corpora than the one used previously, to construct two data sets of comparable size. Our English data set is made of 34 hours and 8 minutes of read speech, and our French dataset contains 33 hours and 42 minutes of read speech. Recordings were sampled at 16000Hz.

We use Kaldi (Povey et al., 2011) to pre-process the data: we extract 13-dimensional MFCCs (25 ms analysis window, 10 ms window shift), to which we apply a vocal tract length normalization (VTLN). We add the Δ and $\Delta\Delta$ for a total of 39 dimensions, and apply centered windowed mean normalization (with a window size of 300 frames).

For each corpus, we use 90% of the data for training and 10% as a validation set. We obtain two models, one for each dataset (**English-DP**, **French-DP**). Model training is stopped after 1500 iterations, as in Chen et al., 2015. We obtain 611 clusters for **English-DP**, and 1565 for **French-DP**.

The **English-DP** and **French-DP** models are applied to the one-second and ten-second test stimuli from the across-speaker condition of the 2017 ZeroSpeech Challenge (also drawn from the LibriVox corpora) and subjected to the corresponding ABX evaluation. We test the French model on the French stimuli and the English model on the English stimuli. The ABX triplets are each made up of a sequence of three extracts of speech from the stimuli, where each extract consists of a sequence of three phones, and A and B differ only in the

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

[5]Catch trials played a tone and gave an audio instruction as to which response to give.

[6]Not all participants used headphones, in spite of our instructions, and a few reported distractions; here we do not exclude these participants. Following a reviewer suggestion, we examined the results of such an exclusion, which leaves 50 English and 26 French participants. All qualitative results remain as reported. The results of this alternate analysis can be found in the online repository.

centre phone, while the context phones are held constant. All triplets constructible from the test stimuli are tested. This test serves to ensure that the models are performing as expected.

We apply each of the two models, separately, to the experimental stimuli (see **Methodology: Human ABX evaluation**), to simulate English and French native listeners. We apply the same pre-processing steps as were applied to the training corpora, transform the files into DPGMM posteriorgrams from the trained models, and obtain only the frames corresponding to the stimuli.[7] We calculate δ for each triplet, for each of the models, and for the MFCC representations.

Results: Humans

The overall ABX discrimination accuracy across all stimuli, across all participants, is 72%. The English listeners obtain a score of 69%, and the French listeners 75%. Figure 1 shows the average accuracy across vowel pairs.[8]

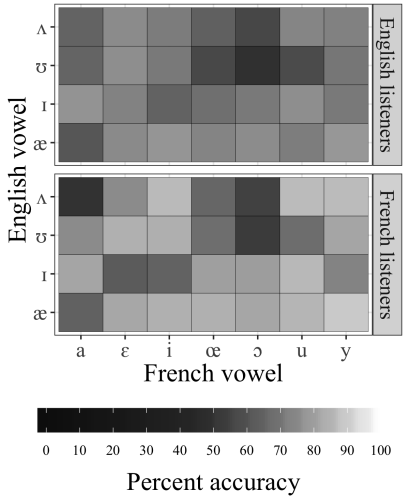

Figure 1: Human accuracy (English and French listeners) averaged by vowel pair. Lighter indicates higher accuracy.

Before comparing the accuracies across native language group, we apply a correction to make the groups' scores comparable. We numerically remove effects of response bias, potential bias to respond A or B, and overall group-level baseline accuracy. We quantify these nuisance effects using a generalized linear model. We fit a probit regression because of its interpretation as a d-prime analysis (DeCarlo, 1998; Macmillan & Creelman, 2004) using the *lme4* package for R (Bates et al., 2015). We code responses as 1 (accurate) or 0 (inaccurate). The model contains an intercept and a random intercept by subject, modelling response bias; a main effect of subject group (English: −1, French: 1), modelling group-level differences; an effect of A/B presentation order (A correct: −1, B correct: 1), modelling tendencies to respond A or B; and an interaction of these last two. We correct each observation by subtracting the predicted probability of correct response. We average the corrected responses within each stimulus triplet, and average these corrected accuracies down to the vowel pair level as before, obtaining corrected accuracies by vowel pair. Correlation between the two groups' corrected accuracy at the stimulus triplet level is 0.63. After averaging to the vowel pair level, the correlation is higher, at 0.79, indicating that many group differences are due to effects of individual stimuli, rather than the vowel contrasts we intended to test. The vowel pairs are compared in Figure 2.

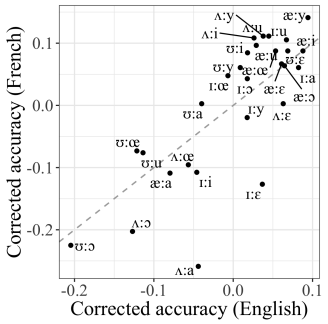

Figure 2: Discriminability of vowel pairs compared between the two language groups. The dotted line is $y = x$; pairs above the line are better discriminated by French listeners, while pairs below show better discrimination for English speakers.

Figure 2 shows that most vowel pairs were relatively well discriminated (upper right), but some were poorly discriminated by both groups (lower left). [ʌ]–[a], [ʌ]–[ɔ], and [ɪ]–[ɛ], are all perceived better by English listeners. This is consistent with Peperkamp (2015), who reports tests of French listeners on identification of English vowels, similarly indicating that, for example, [ʌ] was identified as [a], [œ], or [ɔ].

Results: Models

The scores that **English-DP** and **French-DP** obtain on the ZeroSpeech 2017 stimuli are presented in Table 1. Repeated averaging is done as for the human data, across context (flanking phones), across speakers, and then across all centre phones, to obtain a single score. We observe that the DPGMM model obtains better scores than the MFCCs, consistent with previous results. Results are reported as accuracies. **English-DP** shows 88.4% ABX accuracy on the experimental stimuli we design, and **French-DP** 86.6%, both better than **MFCC** (81.2%). Thus, the models continue to do better, globally, at discriminating speech contrasts, than the acoustic baseline, on novel recordings, from novel speakers.

[7] This was done on the longer source files, rather than directly using the short audio files used in the experiment to avoid window problems, since frames at the beginning and end of files are dropped during preprocessing. Processing the longer source files also gives the vocal-tract length normalization transformation an advantage, leading to an improvement in speaker normalization.

[8] This was a repeated average, similar to that done for the model ABX scores below: first, the accuracy across subjects for a given stimulus was calculated; then, these scores were averaged across contexts; then, across speakers. This was done for consistency with the ABX model evaluation literature (Versteegh et al., 2016; Dunbar et al., 2017).

Model	French		English	
	1s	10s	1s	10s
MFCC	74.8%	74.5%	76.6%	76.6%
French-DP	83.7 %	84.4 %	–	–
English-DP	–	–	88.8%	89.3%

Table 1: ABX accuracy for the trained models and low-level acoustic baseline on the 2017 ZeroSpeech benchmark.

Results: Model–human comparison

To compare the models as models of human perception, we ask how well the continuous machine discriminability score δ for each of the models (distance to incorrect minus distance to correct answer: see **Methodology: Model ABX evaluation**) predicts the human results. As each stimulus is associated with a δ value for a given model, good models are those for which the probability that human subjects respond correctly increases monotonically in the δ value. We compare the three δ values: **English-DP**, **French-DP**, and **MFCC**.

We begin by pooling English and French participants, to assess whether either or both DPGMM models are globally more human-like than the low-level acoustic baseline. We again use probit regression including δ as a predictor. The dependent variable is whether the subject responded correctly (1: accurate, 0: inaccurate). We fit three separate probit regressions, one per δ. Since the model includes a coefficient for δ, this can be seen as taking δ to quantify the subjects' perceived degree of distinctness for a given triplet, up to some scaling factor. We rescale the δ scores for numerical stability and for cross-model interpretability by dividing by the root mean square.[9] We again include both an overall and a (random) by-subject intercept to account for response bias, a coefficient for whether the correct answer was A or B, native language of the participants, and an interaction between these last two, plus a random intercept for individual stimulus triplet (experimental item).[10] We do not include an interaction between subject language and δ: we test for a native language effect separately below. We compare the three models using AIC (Akaike, 1974). Results are in Table 2 (smaller AIC is better). Both DPGMM models predict the human responses better than the MFCC baseline.

If the DPGMM model is really capturing adult perception, we should also expect a "native language effect": the English-trained DPGMM should show results which more closely resemble those of the English listeners than the French listeners, and the French-trained DPGMM should show results which more closely resemble those of the French listeners than the English listeners (see **Results: Humans**). We assess this as follows: we associate each human observation with the appropriate "native language" δ (**English-DP** for trials by English listeners, **French-DP** for French listeners), and with the "non-native language" δ (**French-DP** for English listeners, **English-DP** for French listeners). We construct two alternative probit regression models with the same nuisance predictors as above. In one, the independent variable of interest is the native δ score; in the alternative, the non-native δ. If the representations are equally good at predicting both groups, neither of these models should be better than the other. Results (Table 3) indicate a better fit in AIC for the native-language δ predictor (-21.12 in favour).

To verify that -21.12 is a reasonable model comparison criterion, we examine 9999 instances of the same model comparison over a randomized baseline. Each sample modifies the original data only in that the δ value considered "native" or "non-native" (**English-DP/French-DP**) is determined by a random permutation of the original native language indicator.[11] The random baseline does not yield similar improvements in AIC scores: in the baseline sample, the add-one smoothed left tail probability of -21.12 is 0.0089.

Models	French-DP	English-DP	MFCC
Coefficient for δ	0.2682	0.2790	0.1804
AIC	12675.83	12672.91	12684.15

Table 2: Regressions of human responses against machine representations, compared over the whole experiment (coefficient of δ and AIC). Lower AIC indicates better fit.

Predictor	Native δ	Non-native δ
Coefficient for δ	0.2693	0.1452
AIC	12667.98	12689.1

Table 3: Regressions of human responses against native (**French-DP** for French listeners, **English-DP** for English listeners) versus non-native (switched) trained DPGMM models (coefficient of δ and AIC). Lower AIC indicates better fit.

Discussion

Overall, the DPGMM shows itself to be a passably human-like acoustic model. Furthermore, when it is trained on subjects' native language, it predicts their responses better.

To better understand this effect, we calculate a "degree of native language effect" score for each stimulus triplet in the

[9] We keep zero in place for interpretability, as it is the decision threshold for the model ABX. Note, however, that zero is not guaranteed to be the *optimal* decision threshhold, either for predicting the correct answer in the task, or for predicting human behaviour. The inclusion of an overall intercept allows for the model to adjust to the best decision threshhold for predicting human responses.

[10] We include a stimulus-triplet level random intercept here, but not for the purpose of removing extraneous variability from the accuracy scores in generating Figure 2 above, or Figure 3 below. Those graphs are comparisons of behaviour on different items, and so item-level variability is not a nuisance factor. In contrast, here we are trying to explain away item-level variability, using δ as a predictor. It does not diminish the value of this model comparison to include a predictor capturing additional item-level variability.

[11] By permuting across the data set, we keep the unbalanced proportions of French- and English-native responses. The coefficients for subject language are still fit to the true native language of the subjects.

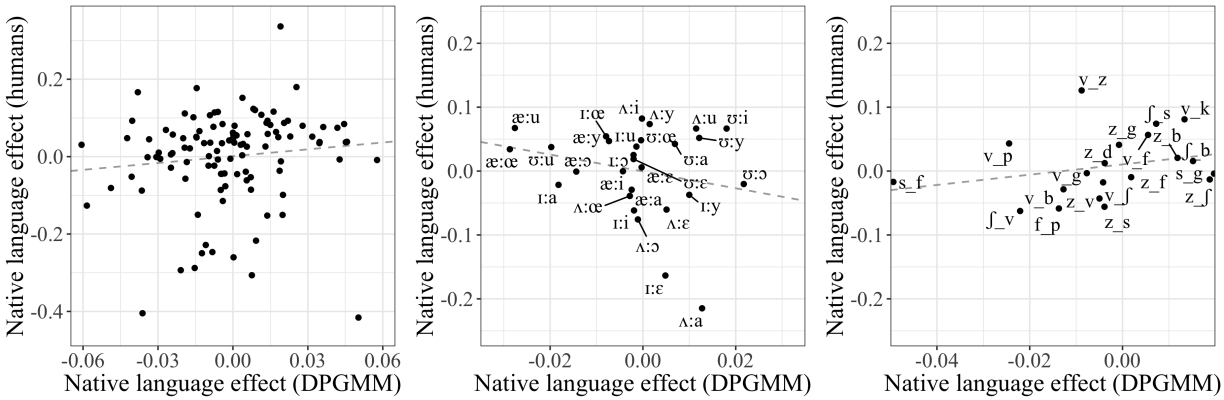

Figure 3: **(a)** Native language effect (French minus English), plotted for human accuracy against probabilities predicted from δ. Each point is one stimulus triplet. **(b)** The same points, averaged by vowel pair. **(c)** The same points, averaged by flanking consonant context. Dotted lines are linear regressions. Graphics do not show the same part of the plane, but all are on the same aspect ratio (13:7), meaning that slopes are visually comparable.

experiment, as the difference between French and English listeners' mean corrected percent accuracy (see **Methodology: Humans**). We calculate an equivalent score for the models, a predicted correct-response probability. Because the mapping between the δ values and response probabilities is indeterminate, we select an optimal mapping: we use a probit regression fit to the human data including the native-language δ as a predictor, and extract the predicted probability for each observation.[12] To isolate the part of the resulting score due to the DPGMM model itself, we subtract from each predicted probability the probability predicted by the regression if δ were zero for the given observation, obtaining a corrected probability analogous to the corrected accuracies derived for the humans above. For each stimulus triplet, we take the average corrected probability across all observations. The native language effect for the DPGMM model, for a given stimulus triplet, is the subtraction of the French and the English models' average corrected probabilities on this triplet.

These quantities are plotted against each other in Figure 3a. The slight trend towards a positive relation is consistent with the results of the model comparison, although most of the variance is unexplained. However, when averaged by vowel contrast, as in Figure 3b, it becomes clear that the native language effect in vowel confusions is not human-like: the trend in the graph is toward a negative relation. Interestingly, in Figure 3c, in which items are instead grouped by consonant frame, shows a slight positive trend, indicating human-like behaviour. But the behaviour the model captures is the fact that the impact of the flanking consonants on performance differs across listener groups. This is clearly not the behaviour we expected it it to capture: the flanking consonants were not intended to have an impact on performance at all. The fact that they contain information that facilitates the task is an artefact of the imperfectly controlled stimuli. It is also not this behaviour that makes the biggest contribution to the native language effect in humans: Figure 3 shows greater variance across vowel pairs than across consonant frames.

This unexpected effect may be due to the nature of the DPGMM model. The large number of categories it learns likely discriminate contextual variants and temporal subcomponents of individual phonemes. The participants in our experiment presumably detect coarser distinctions, beyond this sub-phonemic variability. Vowels, in particular, consist of a long steady state. The DPGMM's representation may fluctuate too much to maintain coarser-grained information. Whatever the explanation, the trained DPGMM models do not match the stimulus-by-stimulus profile of human subjects.

Conclusion

We tested human listeners, English and French native speakers, and an unsupervised acoustic model (trained once on English, once on French) on the same cross-linguistic ABX discrimination task, comparing the model with human performance on a stimulus-by-stimulus level. Our results show that the acoustic model predicts human results better than a low-level acoustic baseline, and predicts certain effects of native language on perception, while missing critical features.

We take this detailed and direct comparison to be an important step in improving the evaluation of quantitative models of human speech perception. Given that the DPGMM shows a limited, but incomplete, correlation with human speech perception, it may also prove useful as a measure of acoustic distance which is adapted to a particular language. Our approach permits detailed investigation of the differences between humans and computational models on speech perception tasks, which will be essential to using these models to gain insight into the underlying cognitive processes.

[12] We use a modified version of the "native language" regression model described in **Results: Model–human comparison**, with all nuisance predictors included, except the random effect of stimulus triplet. We exclude this for reasons discussed already: we are seeking here to examine residual differences between items.

Acknowledgements

This research was supported by the École Doctorale Frontires 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), and ANR-17-EURE-0017.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48.

Chang, J., & Fisher III, J. W. (2013). Parallel sampling of DP mixture models using sub-cluster splits. In *Advances in Neural Information Processing Systems* (pp. 620–628).

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.*

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*(2), 186.

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).

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 1195–1205). Association for Computational Linguistics. Retrieved from `http://aclweb.org/anthology/N18-1108`

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

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Hove, UK: Psychology Press.

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

Peperkamp, S. (2015). Phonology versus phonetics in loanword adaptations. In J. Romero & M. Riera (Eds.), (Vol. 335, pp. 71–90). John Benjamins Publishing Company.

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*.

Schatz, T. (2016). *ABX-discriminability measures and applications*. Doctoral dissertation, École Normale Supérieure.

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.*

Schatz, T., Peddinti, V., Bach, F., Jansen, A., Hermansky, H., & Dupoux, E. (2013). Evaluating speech features with the minimal-pair ABX task: Analysis of the classical MFC/PLP pipeline. In *INTERSPEECH 2013: 14th Annual Conference of the International Speech Communication Association* (pp. 1–5).

Senin, P. (2008). *Dynamic time warping algorithm review.* Retrieved from `http://seninp.github.io/assets/pubs/senin_dtw_litreview_2008.pdf` (Ms., Department of Information and Computer Sciences, University of Hawaii)

Versteegh, M., Anguera, X., Jansen, A., & Dupoux, E. (2016). The Zero Resource Speech Challenge 2015: Proposed approaches and results. *Procedia Computer Science*, *81*, 67–72.

Versteegh, M., Thiollière, R., Schatz, T., Cao, X. N., Anguera, X., Jansen, A., & Dupoux, E. (2015). The Zero Resource Speech Challenge 2015. In *INTERSPEECH-16.*

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245–272.