Pre-registration\_Geomphon\_hindi

Amelia

01/10/2019

# Letter

The process of perception of an unfamiliar speech sound is theorized to involve a comparison of the unfamiliar sound to representations of familiar sounds heard in the past. However, research testing the ability to discriminate an unfamiliar sound often focuses on the difference betweeen the unfamiliar sound and one familiar sound that is determined to be the “nearest” sound in perceptual or acoustic space. This narrow focus on the one nearest sound ignores the rest of a large and often crowded represetation space.

This is important because it is known that sound inventories of language are not random in their geometric shape, but instead can be measured to have global and local symmetry and economy, above and beyond what would be expected by chance. (Dunbar & Dupoux 2016) We ask: does the shape of the entire sound inventory affect discirmination of an unfamiliar sound, above and beyond the effect of distance to the nearest sound? Our research has implications for theories of language change and second language acquisition and practical applications in language teaching.

The study has been approved by the Comité d’Éthique de la Recherche of Université Paris Decartes as of June 13, 2019. All funding for participant payment is covered by the Agence National de Récherche grant numbers ANR-17-CE28-0009 (GEOMPHON) and ANR-10-LABX-0083 (EFL). Web hosting for the online studies will be provided by the Laboratoire de Linguistique Formelle. The study is ready to commence immediately.

Data collection for the English speakers will take a maximum of one week to complete. Based on previous studies we expect recruitment of French participants will take longer, but we anticipate being able to run all participants within a two week window. Once data is collected all code is in place to immediately run models and create vizualizations. At this point we will take one further week to move from a stage 1 manuscript to a stage 2 manuscript. The total time anticipated to go from in principle approval to stage 2 manuscript is one month.

We agree to participate in the Open Peer Review system.

Following Stage 1 in principle acceptance, we agree to register the approved protocol on the Open Science Framework either publicly or under private embargo until submission of the Stage 2 manuscript.

If we should withdraw the paper following in principle acceptance, we agree to the publicaiton of a short summary of the pre-registered study as a Withdrawn Registration.

appropriate reviewers with expertise in linguistics and Bayesian Methods include:

Jeff Mielke  
Timo Roettger  
Bodo Winter  
Till Poppels

# Introduction

### background and literature review

Over a lifetime of speech perception, humans fine-tune their sensitivity to phonetic differences that are contrastive in their native language. When listeners hear a new phone, be it a phone from a new language or a variant they have not previously encountered, they are faced with the task of perceiving new sounds with a perceptual system that is tuned to their native language. There is extensive research investigating precisely how this is accomplished. In formal models of language learning such as the Speech Learning Model (Flege) and the Perceptual Assimilation Model (Best,1995, Best & Tyler 2007), an important concept that has arisen from past work is the distance of the unfamiliar sound to the nearest familiar sound in the native inventory. Unfamiliar sounds will be hard to discriminate from a familiar sound that is close, and easier to discriminate from a familiar sound that is farther away.

This “distance” is defined in different ways: principally, as either an objective acoustic distance measured on some element of an acoustic signal, or as a perceptual distance which can be measured through discrimination tasks. However, a narrow focus on the distance between one familiar sound and one unfamiliar sound ignores the crucial fact that phonetic space is crowded for most sound inventories. An unfamiliar sound which is “far” from one familar sound may be “near” to other sounds, and if a distribution or category develops and changes over time, a move “farther” from one sound is likely to be “nearer” to another sound. We can measure this effect by looking at how the shape of an inventory would change if a new sound were added.

We argue that perception of a phone involves not just the process of comparing the new sound to the nearest old sound, but also as the process of adding new sounds into an already established inventory. In particular, we are interested in how an inventory would change if the new sound is incorporated. It has been shown previously that sound inventories are not random in their distribution of sounds (Dunbar & Dupoux 2016). Dunpar & Doupoux (2016) develop three measurements which encompass these values:

1. economy (“Econ”): if two inventories have the same number of sounds, those sounds tend, all things being equal, to require relatively few features to characterize
2. local symmetry (“Loc”) : if two inventories have the same number of sounds, using the same number of features, those sounds tend to make relatively many “minimal oppositions” between pairs of sounds, which differ only in one feature
3. global symmetry (“Glob”): if two inventories are matched geometrically on the other two properties, they tend to have sounds that are more symmetrically distributed

Because these metrics measure the geometry of an entire inventory, we call these measures “geomphon scores” The calculation of these scores is based on discrete phonological features, and details of the precise method are available in Dunbar & Dupoux (2016). While Dunbar & Dupoux establish that sound inventories of the worlds languages maximize these properties above and beyond the distribution expected by chance, they do not establish *why* inventories maximize these scores. One possible reason is that inventories that maximize these scores are optimal for discrimination of sounds within an inventory. This would lead to the prediction that a change in inventory that reduces one of these scores would lead to worse discrimination, when compared to a change in inventory that increases one of these scores.

In this project, we collect discrimination data from an ABX task in which listeners discriminate pairs of sounds. We use French and English speakers, listening to unfamiliar Hindi sounds.

### Experimental aims

The experiment aims to test whether the spatial relation of all the sounds in a phoneme inventory affect perception of a given unafamiliar sound. Specifically, we test whether changes in geomphon scores are predictive of discrimination accuracy, above and beyond the effect of acoustic distance to one nearest sound.

### Hypotheses

H1) We hypothesize that, as previous research has shown, a pair of sounds that are farther apart in acoustic distance will be easier to discriminate than two sounds that are very close in acoustic distance.

H2) We hypothesize that changes in Econ, Glob, and Loc will each be predictive of discimination accuracy, above and beyond the effect of acoustic distance.

# Method

### sample

#### population

Participants will be recruited in three venues:

1. English speaking participants will be recruited on Amazon Mechanical Turk, with qualifications set to:
   * location is US
   * approval rate 95% or higher
   * have completed > 100 HITS
   * following mechanical turk’s policies, all Amazon Mechanical Turk participants are > 18 years of age
2. French speaking paricipants will be recruited recruited on Amazon Mechanical Turk, with qualifications set to:
   * location is France
   * approval rate 95% or higher
   * have completed > 100 HITS
   * following mechanical turk’s policies, all Amazon Mechanical Turk participants are > 18 years of age
3. French speaking participants will also be recruited through RISC and the facebook groups for linguistics and Psychology research.  
   **DECISION: will we use these? will these speakers be paid?**

#### Inclusion and exclusion criteria

Each participant will answer a demographic survey designed to determine their eligibility. To incentivize truthful behavior in online participants, all Mechanical Turk participants will be paid regardless of their answers.

Data from any subject who meets **any** of the following criteria will be **excluded**.

1. They answer “Yes” to question “Do you have any problems with your hearing or vision, other than glasses/contacts?”
2. They answer “Yes” to question “Have you been diagnosed with a speech or language impairment?”
3. They answer “Yes” to question “Have you ever taken a linguistics course?”
4. **DECISION: include this?? Their answer to the question “Do you have an idea about what we might have been investigating?” indicates that they were aware of the purporse of the experiment.**
5. They indicate they have high proficiency in another language \*\*(They answer \_\_\_ to the question \_\_\_)
6. they indicate they have high experience in another language (They answer \_\_\_ to the question \_\_\_)
7. they indicate they have both medium proficiency and medium experience in another language
8. \*\*DECISION: INCLUDE THIS?Their performance is not significantly different from chance, as determined by\_\_\_\*\*
9. \*\*DECISION: INCLUDE THIS? Their performance is not significantly different from ceiling, as determined by\_\_\_\*\*
10. **DECISION: INCLUDE THIS? Participant has NA/no data for >2 trials (for any reason, e.g.user error, unanticipated technical error)**

We will run a batch of 90 subjects (the total goal number, based on power analysis below), and then run the above analyses to determine the number of subjects that pass all criteria. We will then run batches of size (90 - number who have already passed the above criteria) until we reach 90 subjects. Each batch of data will be posted to OSF in raw, anonymized format on the day of collection, before a new batch is collected.

#### number of subjects

**AMELIA note: (consider adding graph of N based on acoustic distances from disastrous Roger Levy presentation)**

Our goal was to ensure that for our given design there would be enough data that the mean of the posterior distribution would not differ greatly from the true effect. To determine the number of subjects necessary to achive this goal, we tested directly by sampling datasets based on a variety of values of coefficients, and then running models to detect these known coefficient values.

We graph below the difference between the value input as the coefficient and the mean of the posterior for 27 combinations of values (All combinations of 3 coefficient values (-1,0,1), for 3 predictors) . In the best case scenario the mean of the posterior is exactly the same as the coefficient value we used to sample the data, and therefore the difference shown on this graph is zero. We can see that the distribution approaches zero as the number of subjects increases.

An N of *90 subjects* was selected as the best balance between the practicalities of running many subjects and the extent to which the mean of the posterior distribution would reflect the true value.

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## m\_pos\_vars = col\_character(),  
## d\_id = col\_character(),  
## d\_experiment = col\_character(),  
## m\_neg\_vars = col\_character(),  
## is\_correct\_model = col\_logical(),  
## m\_stanfile = col\_character(),  
## csv\_filename = col\_character()  
## )

## See spec(...) for full column specifications.

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## m\_pos\_vars = col\_character(),  
## d\_id = col\_character(),  
## d\_experiment = col\_character(),  
## m\_neg\_vars = col\_character(),  
## is\_correct\_model = col\_logical(),  
## m\_stanfile = col\_character(),  
## csv\_filename = col\_character()  
## )

## See spec(...) for full column specifications.

## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## m\_pos\_vars = col\_character(),  
## d\_id = col\_character(),  
## d\_experiment = col\_character(),  
## m\_neg\_vars = col\_character(),  
## is\_correct\_model = col\_logical(),  
## m\_stanfile = col\_character(),  
## csv\_filename = col\_character()  
## )

## See spec(...) for full column specifications.

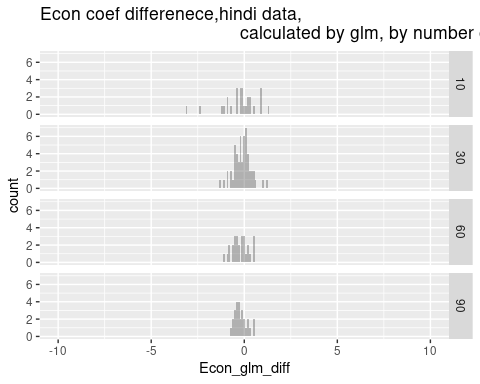
## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## m\_pos\_vars = col\_character(),  
## d\_id = col\_character(),  
## d\_experiment = col\_character(),  
## m\_neg\_vars = col\_character(),  
## is\_correct\_model = col\_logical(),  
## m\_stanfile = col\_character(),  
## csv\_filename = col\_character()  
## )

## See spec(...) for full column specifications.

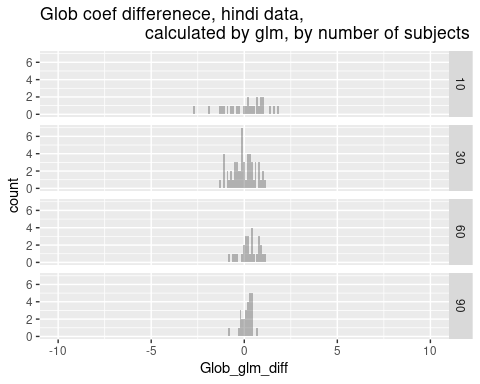
## Parsed with column specification:  
## cols(  
## .default = col\_double(),  
## m\_pos\_vars = col\_character(),  
## d\_id = col\_character(),  
## d\_experiment = col\_character(),  
## m\_neg\_vars = col\_character(),  
## is\_correct\_model = col\_logical(),  
## m\_stanfile = col\_character(),  
## csv\_filename = col\_character()  
## )

## See spec(...) for full column specifications.

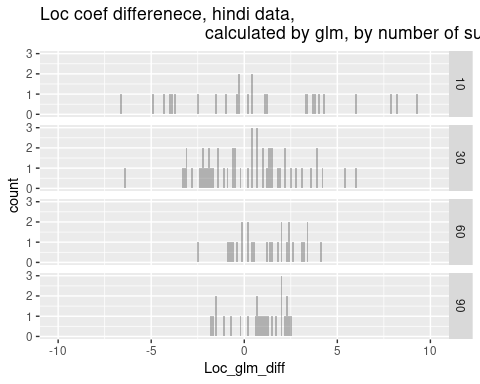
## Warning: Removed 8 rows containing missing values (geom\_bar).



## Warning: Removed 8 rows containing missing values (geom\_bar).



## Warning: Removed 8 rows containing missing values (geom\_bar).



### stimuli

##### Choice of unfamiliar language:

Because this experiment aims to tease apart the effects of acoustic distance and each of the geomphon scores, it was important to find a language that has sounds that are both familiar and unfamiliar. For practical reasons our discrimination experiments are to be run with English and French listeners, and therefore it was necessary to find a language which has some sounds that are similar to sounds in the inventory of French and English, and some sounds that are highly dissimilar. In addition, the unfamilar sounds should involve the best spread of each geomphon score, as well as acoustic distances.

While geomphon scores can be calculated based on a full inventory of consonants and vowels, practically speaking, discrimination between pairs of sounds consisting of one vowel and one consonant would yield ceiling level accuracy and be uninformative for our research question. We therefore use only consonants

Dunbar and Dupoux calculated Geomphon scores for **N** languages and the feature matrices used to do this are available **here** Using these matrices, we identified a list of languages that **criteria**. We then calculated scores base don **fill in details from monday meeting**

Of the languages on this list, Hindi was chosen because it was a language that has many consonants that are familiar to French and English speakers **examples**, and also many that are unfamiliar (in particular, retroflex sounds). It was also a language for which we would be able to recruit speakers in France and the United States. **other reasons we chose Hindi? other details about how this happened?**

##### Procedure for recording of stimuli:

5 Hind-Urdu speakers (3 female) were recorded in a sound attenuated booth in either the Université de Paris or the University of Massachusetts. All reported that Hindi was the principal language they spoke with their family as children. The primary motivation for using multiple speakers was to introduce variability and noise into the experiment, and so the speakers were not controlled for sociolinguistic factors (e.g. age, place of birth,etc.) Each speaker was paid $20 for their recording and signed an informed consent form.

Participants spoke all consonants of Hindi in a pseudoword /a\_a/ context. Each pseudo word was written on a list in Devanagri and latin script, with a sample word in Devanagri including the target consonant sound. For elicitation purposes each pseudoword was said in a sentence context “kripya \_\_\_\_ kahie” **add Devanagri and IPA** (“please bring \_\_\_\_\_”). Speaker 1 recorded all peseudowords once. Speakers 2-5 recorded all consonant sounds twice. The first time, speakers 2-5 listened to speaker 1, and then paused the recording and repeated after Speaker 1. They were told to try to repeat at about the same pace. Recordings are considered individually identifiable information and so are not available publically, but are approved for sharing for research purposes, including replication, by emailing **ewansemail**

Textgrids for all recordings were created using forced alignment **details of tools Nicolas used**. Transcripts, code, and textgrids are available at **URL** All textgrids were hand checked and ajusted for accuracy, and all splice points were moved to the nearest zero crossing.

#### procedure for building and selecting stimuli

Each stimulus tests one pair of phonemes. One member of the pair is a familiar sound, and one member is an unfamiliar sound. It is important to clarify that the “familiar” sounds were hindi sounds spoken by the hindi speakers. However, these were sounds such as /d/ or /z/ that are included in both French and English inventories. This means that the “familiar” sounds may differ systematically from the native versions. However, since these “familiar” sounds are present in the inventories of French and English (i.e. would be transcribed as the same phone in IPA), and the unfamiliar sounds are not, we expect that the familiar sounds are much closer to native sounds (**and indeed our measurements confirm this**). The rationale for performing the experiment this way was to have consistency in the context surrounding the phone of interest. If English speakers recorded the english consonants and Hindi speakers the hindi consonants, the surrounding vowels have the potential to give additional cues useful for discrimination and push discrimination towards ceiling, \*\*as indeed we found in previous pilot\_\_\*\*

The unfamiliar sounds, on the other hand, were sounds which are unambiguously absent in both French and English, such as as /dʒʰ/, /bʰ/, or /ɖ,/.

**stumuli selection section from other rmd goes here**

Once the pairs of phonemes were chosen, it was necessary to determine the order of the phonemes, which sound would be the target, and which speaker’s recording would be used. The number of trials was constrained to 150 based on time estimates of previous experiments. Given the number of permutations of these options it is not feasible to simply counterbalance all combinations of these within the number of trials, which was constrained to 150 trials based on .

Instead, we ensure that the answer is not inferrable from the order or speaker by defining weights for, and use a simulated annealer to optimize a design in which we minimize the extent to which the speaker and order is predictive of the correct answer.

**stimuli are cut out and spliced together LIKE SO**

### calculation of geomphon scores

# **EWAN**

### experimental procedure

Code sufficient to launch a direct replication, including all demographic surveys, instructions, and stimuli, presented in a docker enviroment that replicated exactly the software versions, libraries, and modules we used, can be found here:

\_\_\_PERSISTENT URL.

A description of method is below:

Participants first read a consent form and indicate their consent. Then participants fill out a demographic survey.

Participants hear a triplet

randomization, training, instructions, etc.

### model specifications

#### covariates

**amelia incorporate pre-prereg\_rmd** acoustic distance – how chose how calculated MFCCs? log distance? delta distance?

geomphon scores

#### method to determine whether an effect of a coefficient is non-zero

Based on our design we expect acoustic distance will be non-zero. (indeed, we have specifically designed the stimuli to ensure this is the case)

The effect of the geomphon scores, however, is as yet undetermined, and is expected to be smaller than the effect of acoustic distance *?* Therefore, a significant concern was to determine a model that would detect a difference between a small, positive effect and a zero effect.

To this end, we use comparison of leave-one-out-cross valitation (“loo”) scores with models specified with varying priors. The loo value obtained is a measure of the predictive ability of a model. Specifically, it is an estimate out-of-sample predictive accuracy using within-sample fits. We use the pareto smoothed importance sampling implemented in the loo() package, as documented in Vehtari et al. (2016).

Because we cannot be sure of the size of the expected effect, it is important to establish a methodology that is sensitive to small but non-zero positive effects. To accomplish this we use loo() and truncated priors. A truncated prior is simply a prior for which the density curve is truncated at some value.

In our case, we use a normal(0,10) prior and truncate the value at zero. We can therefore apply a model truncated to be <= 0 to a model to be >= 0. If an effect of interest is not meaningfully different than zero, two models, one with a truncated positive prior and one with a truncated negative prior will not differ (nor would they differ from a model with no effect of this predictor). If, on the other hand, the effect is different from zero, either the truncated positive prior or the truncated negative prior will show better predictivity in the direction of the effect.

For example, to test for an effect of Econ, we will run a model such as the following:

accuracy ~ Econ + Glob + Loc + acoustic\_distance + (1|Subject) + (1|phone\_pair)

we will run the model three times:

* Model A: removing the effect of Econ, (simulating a zero effect)  
  accuracy ~ Glob + Loc + acoustic\_distance + (1|Subject) + (1|phone\_pair),  
  priors: N(0,10) for all covariates
* Model B applying a prior truncated to be >=0  
  accuracy ~ Econ + Glob + Loc + acoustic\_distance + (1|Subject) + (1|phone\_pair),  
  prior Econ: truncated gaussian N(0,10) >=0 all other priors: N(0,10)
* Model C applying a prior trunacted to be <=0 accuracy ~ Econ + Glob + Loc + acoustic\_distance + (1|Subject) + (1|phone\_pair) prior Econ: truncated gaussian N(0,10) <=0 all other priors: N(0,10)

We will then complete pairwise differences to compare the differences in loo scores between each of these four using loo\_compare(), which gives a we used the difference in expected log predictive density (elpd\_diff). The elpd\_diff shows which model has a better loo score, and therefore better predictive accuracy. The output of loo.compare() includes the s.e. of this difference score. We will determine whether two models is meaningfully “different” by using the rule of thumb put forth by Vehtari that the elpd\_diff is 5 times larger than it’s s.e.*?*

In order to make all comparisons of all combinations of the three predictors of interest, we will run a total of 27 models (all combinations of 3 predictors (Econ, Glob, Loc) and three priors (truncated posistive, truncated negative, zero)).

We will then use compare\_models() in the loo package to obtain a ranking of their elpd\_diff scores.

We will interpret there to be a positive effect of one of the geomphon scores Model B is better than model A AND model C.

# Results & Analysis

### Preprocessing

Data will be filtered for whether it confirms to criteria outlined in subejct inclusion/exclusion above. All data that meets criteria will be used in analysis.

*While we have no reason to antifcipate messy data or NAs, Here is our policy*

### reality checks

Our hypotheses take for granted an effect of acoustic distance. Therefore, an effect of acoustic distance is a required first step before interpreting the geomphon scores.(see interpretive plan, below)

### models Any covariates or regressors must be stated.

*Consistent with the guidelines of Simmons et al. (2011), proposed analyses involving covariates must be reported with and without the covariate(s) included.*

All covariates:  
acoustic distance

We will determine whether there is an effe

We will It is not clear on what scale this should be.  
previous pilots show resutls on the order of \_\_\_\_

# Interperative plan

Our interpretive plan therefore will be

step 1: confirm effect of acoustic distance.

possible outcomes:

1. there is no effect of acoustic distance

We will interpret this as a failure to replicate the well-known effect of acoustic distance. Given that the stimuli were chosen specifically to have an effect of acoustic distance, that the number of subjects was chosen specifically to be able to robustly detect this effect, and that three previous pilots show large significant effects of distance, we will we will consider this a fatal failure of the experimental protocol. We will withdraw the study and pre-register a new experimental protocol with more subjects and/ or an easier task in order to detect the effect of acoustic distance.

1. there is an effect of acoustic distance

We will interpret this to be a confirmation that our experimental protocol was well-designed to detect the effects of interest.

step 2: examine the effect of geomphon scores

possible outcomes:

1. there is an effect of one or more of the geomphon features, above an beyond the effect of acoustic distance only.
2. there is no effect of any of the geomphon scores above and beyond the effect of acoustic distance.

# timeline

Timeline for completion of the study and proposed resubmission date if registration review is successful. Extensions to this deadline can be negotiated with the action editor.

## concerns with using loo for comparing multiple models.

Vehtari, Gelman, Gabry 2016: <https://www.arxiv-vanity.com/papers/1507.04544/> “Cross-validation and WAIC should not be used to select a single model among a large number of models due to a selection induced bias as demonstrated, for example, by Piironen and Vehtari (2016).”

Piironen & vehtari 2017 <https://link.springer.com/article/10.1007/s11222-016-9649-y> “Although LOO-CV and WAIC can be used to obtain a nearly unbiased estimate of the predictive ability of a given model, both of these estimates contain a stochastic error term whose variance can be substantial when the dataset is not very large. This variance in the estimate may lead to overfitting in the selection process causing nonoptimal model selection and inducing bias in the performance estimate for the selected model (e.g., Ambroise and McLachlan 2002; Reunanen 2003; Cawley and Talbot 2010). The overfitting in the selection may be negligible if only a few models are being compared but, as we will demonstrate, may become a problem for a larger number of candidate models, such as in variable selection.”

<https://discourse.mc-stan.org/t/interpreting-elpd-diff-loo-package/1628> Vehtari “The difference of 1 SE is definitely too small. If n is not small, and there is no bad observation model misspecification (ie. you have made model checking), and there are no khats>0.7, then as a rule of thumb, I would say that difference of 5 SE starts to be on the safe side.”" " Overall loo is not good for detecting very small differences between models"

<https://avehtari.github.io/masterclass/slides_model_assesment_selection.pdf> slide 151 “Cross-validation can be used for model selection if - small number of models - the difference between models is clear • Do not use cross-validation to choose from a large set of models - selection process leads to overfitting”