Interpreting Asymmetries in Speech Perception with Bayesian Inference

Jie Ren (Jie_Ren@Brown.Edu)
Department of Psychiatry and Human Behavior, Brown University, 1011 Veteran Memorial Pkwy, Riverside, RI 02915 USA

Joseph L. Austerweil (austerweil@wisc.edu)
Department of Psychology, University of Wisconsin – Madison, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

This paper proposes a Bayesian account of asymmetries found in speech perception: In many languages, listeners show greater sensitivity if a non-coronal sound (/b/, /p/, /ɡ/, /k/) is changed to coronal sounds (/d/, /t/) than vice versa. The currently predominant explanation for these asymmetries is that they reflect innate constraints from Universal Grammar. Alternatively, we propose that the asymmetries could simply arise from optimal inference given the statistical properties of different speech categories of the listener’s native language. In the framework of Bayesian inference, we examined two statistical parameters of coronal and non-coronal sounds: frequencies of occurrence and variance in articulation. In the languages in which perceptual asymmetries have been found, coronal sounds are either more frequent or more variable than non-coronal sounds. Given such differences, an ideal observer is more likely to perceive a non-coronal speech signal as a coronal segment than vice versa. Thus, the perceptual asymmetries can be explained as a natural consequence of probabilistic inference. The coronal/non-coronal asymmetry is similar to asymmetries observed in many other cognitive domains. Thus, we argue that it is more parsimonious to explain this asymmetry as one of many similar asymmetries found in cognitive processing, rather than a linguistic-specific, innate constraint.

Keywords: speech perceptual asymmetry; Bayesian inference; natural statistics; category variability; nature vs. nurture; domain-general vs. domain-specific

Introduction

Listeners from a variety of language backgrounds have shown asymmetric sensitivity to different directions of sound changes in speech perception. We focus on one particular asymmetry: consonants with a coronal place of articulation (/d/, /t/, /n/, /l/) are more tolerant to changes or mispronunciations than consonants with non-coronal places of articulation (/ɡ/, /k/, /b/, /p/). For example, the German word for “railway”, bahn, primed the word for “train” when mispronounced as bahn. However, the word for “tree”, baum, did not prime the word for “bush” when mispronounced as baum (Lahiri & van Coillie, 1999). This indicates that listeners can accept a labial sound as the correct form of a coronal sound but not vice versa. ERP findings corroborate this phenomenon: At temporally early stages of speech perception, German-speaking adults displayed asymmetric discrimination for mispronunciations of familiar words with coronal vs. non-coronal onsets (Friedrich, Lahiri, & Eulitz, 2008) and internal consonants (Friedrich, Eulitz, & Lahiri, 2006; Cornell, Lahiri & Eulitz, 2013). Similar effects have also been observed with English-speaking (Roberts, Wetterlin & Lahiri, 2013) and Bengali-speaking adults (Lahiri & Marslen-Wilson, 1991).

To explain this asymmetric bias in speech perception, the predominant hypotheses have been derived from phonological theories (Kiparsky, 1982) in the framework of Universal Grammar (UG). In particular, the Featurally Underspecified Lexicon (FUL) model (Lahiri & Marslen-Wilson, 1991; Lahiri & Reetz, 2002) suggested that the place of articulation for coronal consonants (/d/ in duck) is not stored (underspecified) in phonological representations. Consequently, mispronunciations in the onset of duck, such as guck, are still compatible with the representation of duck, and such mispronunciations should have minimal effects on the lexical activation of the word duck. By contrast, the place of articulation for a non-coronal consonant is fully stored (specified) in the phonological representation (the place of articulation of /ɡ/ in goose is stored as [+velar]), so mispronunciations of the onset of goose, such as doose, will be incompatible with the representation of goose and thus will disrupt lexical activation of the word goose.

Studies with infants and toddlers also support predictions of the FUL model. 6-month-old Dutch-learning infants were habituated to repeated taan or paan tokens and then tested on their ability to discriminate trials in which the stimulus repeated versus trials in which the two stimuli alternated. Whereas infants habituated to paan discriminated the two types of trials, infants habituated to taan did not (Dijkstra & Fikkert, 2011). The findings were interpreted as support for the FUL model: When the standard of comparison was taan, the place of articulation of the onset /t/ was not specified, so both taan and paan were compatible with the standard. But when the standard of comparison was paan, place of articulation of the onset /p/ was specified, so only paan was compatible with the standard, and paan and taan were discriminable. 4- and 6-month-old Dutch- and Japanese-learning infants were also tested using the same procedure on their discrimination of labial (omba) and coronal (onta) sounds (Tsui, Mazuka, Christia, & Fikkert, 2015). Infants habituated to the labial sound omba discriminated the two types of trials, but infants habituated to the coronal sound onta did not, regardless of their language background.

Linguistic hypotheses, such as underspecification, provide one source of explanations for the observed asymmetry in speech perception. Consistent with the FUL model, Fennel
(2007) showed that infants detected a labial-to-coronal switch but failed to detect a coronal-to-labial switch. However, inconsistent with this hypothesis, a follow-up study (Fennell, van der Feest, & Spring, 2010) showed that 14-month-olds were better able to detect a coronal-to-velar switch than a velar-to-coronal switch. To explain such discrepancy, they analyzed the onset formant frequencies of all the /b/, /d/ and /g/ tokens in their experimental stimuli, and discovered that /b/</d/</g/ in acoustic variability. Thus, the authors concluded that the asymmetries they observed might be better explained by acoustic variability than phonological specification.

We examine the same asymmetries in speech perception in the framework of Bayesian inference: Asymmetric perception of coronal and non-coronal places of articulation may arise from differences in the statistical properties of the coronal category and the non-coronal category in competition (e.g. within word minimal pairs). As presented later, coronal consonants as a category are more frequent and/or variable than non-coronal consonants in languages where asymmetric perceptions have been found. These statistical properties yield an asymmetric posterior distribution: Given a speech signal equidistant from a prototypical coronal and noncoronal articulation, the signal is more likely to be a coronal consonant.

**The Model**

**Theoretical Overview**

Following the tradition of categorical perception of speech sounds by Liberman et. al. (1957), Clayard et. al. (2008) and Feldman, Griffiths, and Morgan (2009), we interpret the asymmetry as the result of statistical inference of speech categories from a noisy speech signal. Listeners utilize available information from a variety of sources to achieve such a goal, including their prior knowledge of native speech categories and the acoustic properties of the speech signal.

A speech category is defined in the model as a distribution over acoustic dimensions. According to the model, when speakers articulate a sound, they first choose a speech category and then articulate a sound exemplar from that category. Each sound exemplar of the speech category varies from one another due to many factors, such as coarticulation, affect, focus and grammatical intonation. Although speaker’s articulation over acoustic dimensions is multidimensional, for mathematical simplicity we assume articulations of a speech category can be reduced to a Gaussian distribution over a single acoustic dimension. Thus, the inventory of native speech categories is represented as a set of Gaussian distributions in the model. Different speech categories differ in the location of their means and in how much their articulation varies over the acoustic dimension (variance). In addition, categories may differ in frequency of occurrence with some categories used more frequently than others. The frequency of occurrence of each category is represented by its prior probability.

Listeners assume that the perceived signal was generated by a speech category that is masked by noise, including environmental noise and perceptual errors. The listeners’ task is to recover the speech category that is most likely to have produced the speech signal. If there are two categories that could have generated the speech signal, listeners should take into account of both categories by weighing the statistical properties of each category. Suppose that in a hypothetical language, coronal and non-coronal categories have equal variance and equal frequency of occurrence. Then each of the two categories has an equal posterior probability to have generated a speech signal equidistant from the mean of the coronal and non-coronal distributions. However, in real languages, the coronal category is often higher in variances and/or frequencies of occurrence (the tip and blade of the tongue are more flexible and more variable). An ideal observer should take these factors into account, which may result in the posterior probability of the equidistant speech signal to be larger for the coronal than the non-coronal category.

**Mathematical Formulation**

Here we formulate a Bayesian model of speech perception. Although we apply the model to the asymmetric perception of coronal and non-coronal consonants, the model may apply to any domain where a person observes a noisy signal from categories, with each category’s exemplars being approximately Gaussian distributed.

We consider speech perception as probabilistic inference. Listeners infer the category membership $C_i$ of a noisy signal $S$, as denoted by the conditional probability $p(S \mid C_i)$. We denote $i = 1$ for coronal membership and $i = 2$ for non-coronal membership. The posterior probability $p(C_i \mid S)$ that an observed noisy signal $S$ is a coronal sound can be obtained by Bayes’ Rule:

$$p(C_i \mid S) = \frac{p(S \mid C_i) p(C_i)}{p(S \mid C_1) p(C_1) + p(S \mid C_2) p(C_2)}$$

(1)

$p(C_i)$ in Equation 1 is the prior probability of the coronal category and $p(S \mid C_i)$ is the likelihood of observing stimulus $S$ given it was generated by a coronal category.

Now we derive a closed form solution to the posterior probability that a signal $S$ is coronal, $p(C_i \mid S)$. Suppose that the speaker articulates an exemplar $E$ of the coronal category $C_i$ and $E$ is Gaussian distributed with mean $\mu_{c_i}$, the prototype of the coronal category $C_i$. Exemplars within a category vary with variance $\sigma_{c_i}$. Therefore,

$$E \mid C = C_i \sim N(\mu_{c_i}, \sigma_{c_i}^2)$$

The speaker’s articulation, the speech signal passing through the environment, and the perceptual system of the listener all add noise to the exemplar. These sources of noise

---

1 The model may also account for other asymmetries in speech production. We plan to pursue this in future work.
combined into $\sigma_s^2$. Therefore, the conditional distribution of $S \mid E$ is

$$S \mid E \sim N(E, \sigma_s^2)$$

where $\sigma_s^2$ represents the random noise that is not due to within-category variability $\sigma_c^2$. Due to conjugacy, $E$ can be marginalized out to form the likelihood $p(S \mid C_i)$, which is Gaussian distributed:

$$S \mid C_i \sim N(\mu_{c_i}, \sigma_c^2 + \sigma_s^2)$$  \hspace{1cm} (2)

The likelihood’s variance is the sum of two components: the category variance $\sigma_c^2$, and random, environmental, and perceptual noise $\sigma_s^2$. Plugging in the parameter values into a Normal distribution, Equation 2 can be written as:

$$p(S \mid C_i) = \frac{1}{2\pi \sigma_c^2 + \sigma_s^2} \exp \left\{ -\frac{(S - \mu_{c_i})^2}{2(\sigma_c^2 + \sigma_s^2)} \right\}$$  \hspace{1cm} (3)

Following the same logic, the likelihood of the non-coronal category $p(S \mid C_2)$ is

$$p(S \mid C_2) = \frac{1}{2\pi \sigma_c^2 + \sigma_s^2} \exp \left\{ -\frac{(S - \mu_{c_2})^2}{2(\sigma_c^2 + \sigma_s^2)} \right\}$$  \hspace{1cm} (4)

Plugging Equations 3 and 4 into Bayes Rule (Equation 1), we can rewrite the posterior probability of the coronal category given the perceived speech signal $S$ as

$$P(C_1 \mid S) = \frac{1}{1 + \beta_1 \exp(\beta_2 S^2 + \beta_3 S + \theta)} \frac{P(C_2)}{P(C_1)}$$  \hspace{1cm} (5)

where

$$\beta_1 = \frac{\sqrt{\sigma_c^2 + \sigma_s^2}}{\sqrt{\sigma_c^2 + \sigma_s^2}}$$

$$\beta_2 = \frac{(\sigma_c^2 - \sigma_s^2)}{2(\sigma_c^2 + \sigma_s^2)}$$

$$\beta_3 = \frac{2(\mu_{c_2} - \mu_{c_1})}{2(\sigma_c^2 + \sigma_s^2) * (\sigma_c^2 + \sigma_s^2)}$$

The closed form solution for the posterior is given by Equation 5. We explore how the relative differences in variability and frequency between coronals and non-coronals affect the posterior probability of coronals $P(C_1 \mid S)$. Then, we analyze the natural statistics of coronals and non-coronals to determine whether perceptual asymmetries would arise from them in an ideal observer.

**Quantitative Evaluation**

Suppose that in a hypothetical language, the coronal category and the non-coronal category are equally frequent – have equal priors, $P(C_1) = P(C_2)$. Also suppose that the categories are equally variable, as encoded by $\sigma_c^2 = \sigma_c^2$. In these circumstances, Figure 1 depicts the posterior probability for a noisy speech signal to be perceived as a coronal sound, $P(C_1 \mid S)$.

Given equal variance and equal frequency of occurrence, the category boundary divides the perceptual space into two equal parts. This indicates that a noisy signal equidistant from the category prototypes has an equal probability of being perceived as a coronal or a non-coronal. We now examine how heterogeneity of category variances (i.e. if $\sigma_c^2 \neq \sigma_c^2$) and unequal frequency (i.e., if $P(C_1) \neq P(C_2)$) affect the posterior probability of the coronal category $P(C_1 \mid S)$.

**Effect of Category Variance**

In many languages where perceptual asymmetries are found, exemplars of the coronal category are more variable than exemplars of the non-coronal category. For example, using a modified Levenshtein distance metric, Cohen-Priva (2012) aligned the underlying (dictionary) forms and phonetic realizations in the Buckeye Natural Speech Corpus (Pitt, Johnson, Raymond, Hume & Fosler-Lussier, 2007). He created an articulatory confusion matrix for English segments in the corpus. Of the 43,915 coronal stop tokens, 21,576 (49%) were pronounced either as allophonic variants or as some other phonemes, whereas of the 64,288 noncoronal stop tokens, only 2,997 (5%) were pronounced as allophonic variants or as an alternative phoneme. Such analyses indicate that coronal stops are about 10 times more variable than noncoronal stops. Moreover, coronals (9% of all coronal segments; 20% of coronal stops were deleted) were also more likely to be deleted than noncoronals (5% of all noncoronal segments, 4% of noncoronal stops were deleted).

The differences in natural language statistics of the within-category variances between coronal and non-coronal categories in the Buckeye corpus provide the following constraint: $\sigma_c^2 > \sigma_c^2$. Suppose that $\sigma_c^2 = 5\sigma_c^2$ (approximately the difference in the segment deletion rates for the English data in the Buckeye corpus), the posterior probability for a noisy speech signal to be perceived as a coronal sound is displayed in Figure 2 (with the case where the within-category variances are equal for reference).

As the red dashed curve shows in Figure 2, the category boundary has shifted towards the non-coronal category, leaving a larger posterior probability for a noisy signal equidistant between the categories to be perceived as a coronal sound. Suppose that the posterior probability for the speech signal $[g]$ to be perceived as a coronal sound $/d/$ is 0.1 given equal variance (blue curve). The shift of category boundary (dashed red line) leads to an increase in the posterior probability for $[g]$ to be perceived as $/d/$ (to a value around 0.2). Thus, due to the higher variance of the coronal
category, an ideal listener is more likely to perceive /g/ as an exemplar of the /d/ category. Now we examine the reversed direction, i.e. a coronal ([d]) signal is changed to a non-coronal sound (/g/) given unequal variance.

Coronal segments in these languages are at least twice as frequent as either labial or velar segments (Japanese velar stops are exceptional and we will discuss this case later).

Frequency is represented by prior probabilities in the model. \(P(C_1)\) and \(P(C_2)\) are the prior probabilities of the coronal category and the non-coronal category, respectively. Suppose that \(P(C_1) = 2P(C_2)\) (approximately the relative frequency in Table 1).

![Figure 2: Posterior probability (category boundary) of the coronal category as a result of unequal variance.](image)

![Figure 3: Posterior probability (category boundary) of the non-coronal category as a result of unequal variance.](image)

Table 1: Frequencies of segments in CALLHOME transcripts by place of articulation

<table>
<thead>
<tr>
<th>Language</th>
<th>Consonant Segments</th>
<th>Labial</th>
<th>Coronal</th>
<th>Velar</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic</td>
<td>All</td>
<td>91,409</td>
<td>222,774</td>
<td>94,624</td>
</tr>
<tr>
<td></td>
<td>Stops</td>
<td>25,592</td>
<td>54,279</td>
<td>38,544</td>
</tr>
<tr>
<td>Japanese</td>
<td>All</td>
<td>57,513</td>
<td>236,813</td>
<td>99,760</td>
</tr>
<tr>
<td></td>
<td>Stops</td>
<td>15,854</td>
<td>62,241</td>
<td>79,117</td>
</tr>
<tr>
<td>Spanish</td>
<td>All</td>
<td>101,717</td>
<td>320,167</td>
<td>53,483</td>
</tr>
<tr>
<td></td>
<td>Stops</td>
<td>44,366</td>
<td>62,961</td>
<td>43,005</td>
</tr>
</tbody>
</table>

Figures 4 and 5 show the posterior probability of the coronal category \(P(C_1|S)\) and the non-coronal category \(P(C_2|S)\), respectively. The category boundary (green dashed curve) has shifted towards the non-coronal category due to the larger prior probability of the coronal category. For comparison, the posterior probability given equal prior probabilities of coronals and non-coronals is plotted as the blue curve. This results in a larger posterior probability for a noisy non-coronal signal ([g]) to be perceived as a coronal sound (/d/) (Figure 4) and a smaller posterior probability for a noisy coronal signal ([d]) to be perceived as a non-coronal sound (/g/) (Figure 5).

![Figure 4: Posterior probability (category boundary) of the coronal category as a result of unequal frequency.](image)

![Figure 5: Posterior probability (category boundary) of the non-coronal category as a result of unequal frequency.](image)
Thus, similar to the effect of variance, larger frequency of occurrence of the coronal category also causes an ideal listener to be more likely to accept that the noisy signal [g] is an exemplar of the /d/ category but not vice versa, predicting the same pattern of asymmetry.

Japanese velar sounds provides an interesting test for an ideal listener model. As Table 1 shows, velar stops (/k/ and /kʰ/) occur more often than coronal stops (/t/) in Japanese. Our model predicts that assuming equal variance, the pattern of asymmetries should be reversed for Japanese listeners—they should be less sensitive to sound changes from non-coronal to coronal than vice versa. Japanese studies (see Tsuji et. al, 2015) so far have only tested infant listeners with labial (omba) and coronal (onta) but not velar (/k/ and /kʰ/) phonemes. Future experimental studies should examine this prediction with these velar and coronal phonemes.

**Effect of Variance and Frequency** In many languages both the prior distribution and the variance of the coronal category are larger than those of the non-coronal category. Here we examine how prior and variance interact.

![Figure 6: Posterior probabilities of the coronal category as a result of unequal frequency and/or unequal variance](image)

**Figure 6: Posterior probabilities of the coronal category as a result of unequal frequency and/or unequal variance**

![Figure 7: Posterior probabilities of the non-coronal category as a result of unequal frequency and/or unequal variance](image)

**Figure 7: Posterior probabilities of the non-coronal category as a result of unequal frequency and/or unequal variance**

Figure 6 shows the posterior probabilities of acoustic signals to be perceived as a coronal sound and Figure 7 shows the posterior probabilities of acoustic signals to be perceived as a non-coronal sound with differing assumptions regarding the relative frequency and variance of coronal and non-coronal sounds. The grey line shows category boundary shift as the result of both the larger prior and variance of the coronal category. As shown in both figures, the pattern of asymmetry remains the same, but there is an even larger posterior probability for a non-coronal signal to be perceived as a coronal (Figure 6), and an even smaller posterior probability obtained for a non-coronal signal to be perceived as a coronal sound (Figure 7). Thus, with larger variance and larger frequency of occurrence, an ideal listener is even more likely to perceive a non-coronal signal ([g]) as a coronal sound (/d/), but even less likely to perceive a coronal signal ([d]) as a non-coronal sound (/g/).

**General Discussion**

We presented an alternative account for the asymmetry in perceiving coronal and non-coronal consonants in speech processing: They arise due to Bayesian inference given the natural statistics of coronals and non-coronals. Listeners make use of their represented category frequency and variance to make inference about the category membership of a perceived signal. Asymmetry occurs when the two speech categories in competition (e.g. within a word minimal pair) are not equal in variance and/or frequency of occurrence.

Our approach diverges from the currently predominant approach in linguistics, which explains the asymmetric perception as due to underspecification of the coronal place of articulation. This theory relies on the special phonological status of coronal sounds only. Conversely, our approach accounts for the asymmetry as due to the relative statistical properties of different speech categories. The underspecification hypothesis is a language-specific, innate constraint, whereas our account is experience- (learning-) based and domain-general. For example, Quinn, Eimas & Rosencrantz (1993) found that 4-month-olds habituated to pictures of cat faces could easily detect a change to a picture of a dog face. However, infants habituated to dog faces failed to detect a change to a cat face. A series of follow-up studies investigating this asymmetry confirmed that dog stimuli are more variable in appearance and that when variability was equated across categories the asymmetry disappeared (Eimas, Quinn & Cowan, 1994; Mareschal, French & Quinn, 2000). In music perception, Delbé, Bigand & French (2006) examined effects of variability by training non-musician adults with two distributions of pitch sequences differing in variability and then testing them on sensitivities to the two directions of changes. Results indicated that changes from the less variable category to the more variable category are more detectable than vice versa.

Our account derives predictions to test in future work. First, category frequencies and within-category variances are learned from early language exposure. In languages where non-coronals are more frequent and/or vary more within-category, the model predicts that the asymmetry should be reversed. Second, at any developmental stage when listeners have stable representations of the corresponding frequency of occurrence and variance of two competing categories, perceptual asymmetries may occur in speech processing. Third, frequency and variance should have independent effects on speech processing. Further, for mathematical simplicity we assumed that a speech category is a Gaussian distribution over a single acoustic dimension.
Thus, the model does not differentiate between different sources/dimensions of variability (e.g., contextual effect, Ganong, 1980). Assuming we can control these factors experimentally and test the posterior probability of coronal and non-coronals in a fine-grained manner, the model makes quantitative predictions as to the precise form of the asymmetry. None of these predictions arises from UG but from the statistics of speech input exposure.

It is also worth noting that not all asymmetries in speech or other cognitive domains are caused by category natural statistics. Findings on vowel (Polka & Bohn, 2003) and face (Corneille, Goldstone, Queller & Potter, 2006) perception, for example, have suggested that similar asymmetric patterns could also be due to stimulus saliency and experimental training. Future studies may examine how these factors interact with frequency of occurrence and variances in category perception.

In conclusion, we have provided a novel explanation for the asymmetry between coronal and non-coronal sounds in speech perception. Whereas phonological specification as a hypothesis could be useful for linguistic purposes, it is not necessary to account for asymmetries in speech perception.

References


