The effect of pre-linguistic normalization in vowel perception

Anna Persson[1] & T. Florian Jaeger[2,3]

[1] Swedish Language and Multilingualism, Stockholm University

[2] Brain and Cognitive Sciences, University of Rochester

[3] Computer Science, University of Rochester

Author Note

We are grateful to ### ommitted for review ###

Correspondence concerning this article should be addressed to Anna Persson, Department of Swedish Language and Multilingualism, Stockholm University, SE-106 91 Stockholm, Sweden. E-mail: anna.persson@su.se

Abstract

XXX. All data and code for this study are shared via OSF, including the R markdown document that this article is generated from, and an R library that implements the models we present.

Keywords: speech perception; vowels; distributional learning; computational model

Word count: XXX

The effect of pre-linguistic normalization in vowel perception

1 To Do

- check that treatment coding for glm is correct (7 variables instead of 3)

2 Abstract

One of the central challenges for speech perception is that talkers differ in pronunciation–i.e., how they map linguistic categories and meanings onto the acoustic signal (Liberman et al., 1967). While this challenge is always present, it is most evident when listeners first encounter talkers with unfamiliar pronunciations. Yet, listeners typically overcome even these difficulties within minutes (e.g., Clarke & Garrett, 2004; Xie et al., 2017). What mechanisms underlie these adaptive abilities remains unclear. One highly influential general hypothesis holds that inter-talker differences are removed via low-level pre-linguistic auditory normalization of acoustic cues. There are now at least a dozen of competing normalization proposals (e.g., Lobanov, 1971; Nearey, 1978; McMurray & Jongman, 2011). Despite the fact that these proposals differ in the cognitive capacities they entail, comparisons of different normalization models against human perception remain largely lacking (but see Bion & Escudero, 2007; McMurray & Jongman, 2011). Here, we seek to address this gap by comparing normalization accounts against the perception of American English vowels.

We first trained 7 ideal observer (IO) models that differed in whether they were trained on unnormalized or normalized acoustic cues. The following influential normalization models was employed: C-CuRE, Lobanov, Miller, Gerstman, Nearey1 and Nearey2. All models were trained against the same phonetic database of productions of all 8 h-VOWEL-d words of American English (heed, hid, head, had, odd, hut, hood, who'd, N=9 tokens per vowel from 17 talkers each), previously described in (REF to Li & Xie). We then compared the predictions of all IOs against L1 American English listeners' 8-way categorization responses for productions of the 8 h-VOWEL-d words in a web-based experiment (N=22 participants).

The best performing IO only centered cues relative to talkers' cue mean (as in C-CuRE normalization, mean accuracy 64.5%, SE 1.0%). This is significantly above chance (12.5%) but

also significantly below the best possible performance (always guessing the response most frequently given by human listeners, 72.0%). Statistically indistinguishable performance was achieved when cues were both centered and scaled (as in, e.g., Lobanov normalization, 63.5%, SE 1.0%). This suggests that simple normalization operations might be sufficient to explain perception. Either type of normalization model performed significantly better than IOs based on unnormalized cues (53.2%, SE 0.7%). Additional comparisons showed that the benefit of normalization over unnormalized cues held regardless of whether cues were first transformed from acoustic (Hz) into perceptual spaces (Mel, Bark, and similar). In conclusion, these results indicate that pre-linguistic normalization (or computationally similar algorithms) contribute to the remarkable adaptive abilities of human speech perception. Our results further suggest that human perception only employs simple normalization operations—such as centering cues relative to a talker's mean. Finally, we find that simple ideal observers achieve performance far above chance in predicting perception, although our results also indicate that human perception might employ additional adaptive algorithms (given the 7.5% accuracy difference between the best performing models and human performance).

3 Introduction

One of the central challenges for speech perception is that talkers differ in pronunciation–i.e., how they map linguistic categories and meanings onto the acoustic signal (Liberman et al., 1967). While this challenge is always present, it is most evident when listeners first encounter talkers with unfamiliar pronunciations. Yet, listeners typically overcome even these difficulties within minutes (e.g., Clarke & Garrett, 2004; Xie et al., 2017). What mechanisms underlie these adaptive abilities remains unclear. One highly influential general hypothesis holds that inter-talker differences are removed via low-level pre-linguistic auditory normalization of acoustic cues. There are now at least a dozen of competing normalization proposals (e.g., Lobanov, 1971; Nearey, 1978; McMurray & Jongman, 2011), that have indeed been shown to reduce irrelevant inter-talkers variability due to e.g. anatomical or physiological factors. Despite the fact that these proposals differ in the cognitive capacities they entail, comparisons of different normalization models against human perception remain largely lacking (but see Bion & Escudero, 2007;

McMurray & Jongman, 2011). Here, we seek to address this gap by comparing normalization accounts against the perception of American English vowels.

Normalization procedures that reduce the category variance and increase category separability should generally reduce perceptual difficulties, hence improve speech perception/categorization. Reducing category overlap is known to increase category separability in perception (Feldman, Griffiths, and Morgan (2009); Kleinschmidt and Jaeger (2015); Kronrod, Coppess, and Feldman (2016)). Following previous research, we hypothesize that normalization procedures that have previously been shown to reduce inter-talker variability, would better explain human perception of vowels… The following influential normalization models was employed: C-CuRE, Lobanov, Miller, Gerstman, Nearey1 and Nearey2.

4 Methods

In order to assess the effect of normalization procedures, we use a computational model based on Bayesian probability theory, ideal observers (see e.g. Feldman et al., 2009; Kleinschmidt & Jaeger, 2015; Kronrod et al., 2016). This allows us to simulate the effects of normalization procedures and to ask what the predicted consequences are for perception. We trained 7 ideal observer (IO) models on unnormalized or normalized acoustic cues from a phonetic database of American English, previously described in (REF to Li & Xie). The Li & Xie corpus consists of recordings from 15 native (five female) and 15 non-native (five female) talkers of a Northeastern dialect of American English (ages 18 to 35 years old). For this study, we selected the male and female native talkers (N=17). The IOs were trained on all 8 h-VOWEL-d words of English in the database (heed, hid, head, had, odd, hut, hood, who'd, N=9 tokens per vowel from 17 talkers each). The IOs' categorization performances were compared against each other, as a measurement of the efficiency of each model in reducing irrelevant inter-talker variability in the data. The predictions of the IOs were then evaluated against human categorization data from a web-based experiment on vowel categorization of English vowels.

4.1 Vowel categorization experiment

In order to evaluate the IOs' predictions against human perceptual data, we exposed native English listeners (N=22) to the all vowel productions from one of the female native English talkers in the phonetic database (9 repetitions * 8 vowels = 72 words) in an 8-way categorization experiment. The experiment was administrated on Amazon Mechanical Turk and consisted of one test block with 144 trials (each word repeated twice over the experiment). Participants were instructed to listen to a female talker saying words, and click on a word on screen to report what word they heard. At each trial, all eight hVd-words were displayed on screen in a response grid. Eligibility requirements, besides being a native speaker of American English, were to complete the experiment in a quiet room, wearing over-the-ear headphones of good sound quality. Before taking the experiment, participants performed a sound check and signed a consent form. After completing the experiment, participants filled out a language background questionnaire and were reimbursed.

4.2 Normalization procedures

The normalization procedures used in this study are listed in Table 1. We selected procedures that are commonly used in speech production and perception research and that have been compared and evaluated in several studies (see e.g. (**adankComparisonVowelNormalization2004a?**), XXX). All procedures take formant frequencies (in Hertz) for the first two formants, F1 and F2, as input and outputs normalized versions of the same formants.

Table 1
Normalization procedures selected

Normalization procedure	Source
Miller (formant-ratio)	(Miller, 1989)
Nearey1 (logmean)	(Nearey, 1977)
Nearey2 (shared logmean)	(Nearey, 1977)
C-CuRE (centering; subtracting from expected)	(McMurray et al., 2011)
Lobanov (z-score)	(Lobanov, 1971)
Gerstman (range normalization)	(Gerstman, 1968)

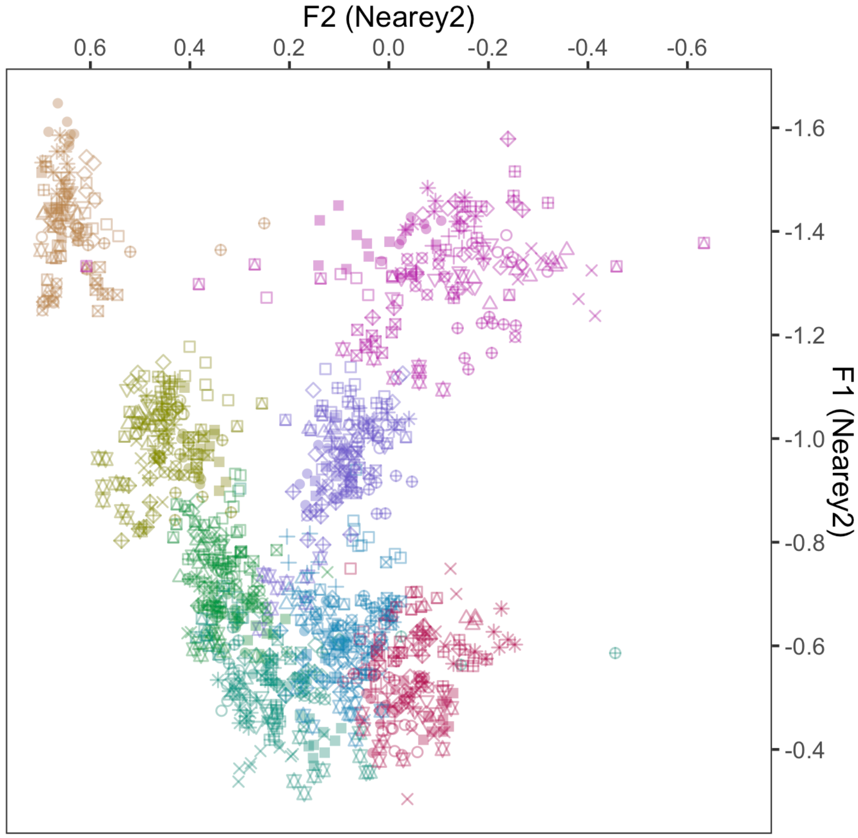
4.2.1 Training-test data split

Prior to modelling the effects of different normalization procedures on the perception of English vowels, we split the vowel data into training and testing portions in a K-fold cross-validation approach in order to minimize the risk of overfitting the models. Overfitting would in this case mean that the ideal observers learn overly specific vowel categorization functions that perform well on the vowel data they have been trained on, but cannot generalize to unseen data. We set K = 5, and split the vowel data randomly into five equally sized bins. We then trained five ideal observers on four of the bins in a latin square design, and tested each ideal observers' predictions for vowel perception on the test data from the vowel categorization experiment.

```
Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
i Please use `all_of()` or `any_of()` instead.
  # Was:
  data %>% select(cues)

  # Now:
  data %>% select(all_of(cues))

See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

The test data from the perceptual experiment is then subsequently normalized with the same normalization procedures, but using the parameters generated from running the procedures on the training data. This means that we use the means and variances obtained from the training data to normalize the test data for all normalization procedures that make use of formant means and variances. This applies to Lobanov normalization, Nearey1, Nearey2, Miller, C-CuRE and Gerstman?????? (min,max). MOTIVATION ———————————— simulating that learning has finished when exposure is over???

```
Warning: Removed 36 rows containing missing values (`geom_point()`).
```



5 Results

The best performing IO only centered cues relative to talkers' cue mean (as in C-CuRE normalization, mean accuracy 64.5%, SE 1.0%). This is significantly above chance (12.5%) but also significantly below the best possible performance (always guessing the response most frequently given by human listeners, 72.0%). Statistically indistinguishable performance was achieved when cues were both centered and scaled (as in, e.g., Lobanov normalization, 63.5%, SE 1.0%). This suggests that simple normalization operations might be sufficient to explain perception. Either type of normalization model performed significantly better than IOs based on unnormalized cues (53.2%, SE 0.7%). Additional comparisons showed that the benefit of normalization over unnormalized cues held regardless of whether cues were first transformed from acoustic (Hz) into perceptual spaces (Mel, Bark, and similar).

5.1 Predicting the category heard by listeners

We tested the IOs in predicting the category heard by listeners in the categorization experiment.

```
Warning: Using an external vector in selections was deprecated in tidyselect 1.1.0.
i Please use `all_of()` or `any_of()` instead.
  # Was:
  data %>% select(levels.vowel.IPA)

  # Now:
  data %>% select(all_of(levels.vowel.IPA))

See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
```

Figure **??** visualizes the accuracy of the seven IOs in predicting the vowel responses from the 22 native English participants in the vowel categorization experiment.

```
Warning: `cols` is now required when using `unnest()`.
i Please use `cols = c(ci)`.
`cols` is now required when using `unnest()`.
i Please use `cols = c(ci)`.
```

```
$predicates
<list_of<quosure>>

[[1]]
<quosure>
expr: ^IO.cue_normalization == "C-CuRE (Hz)"
env:  global
```

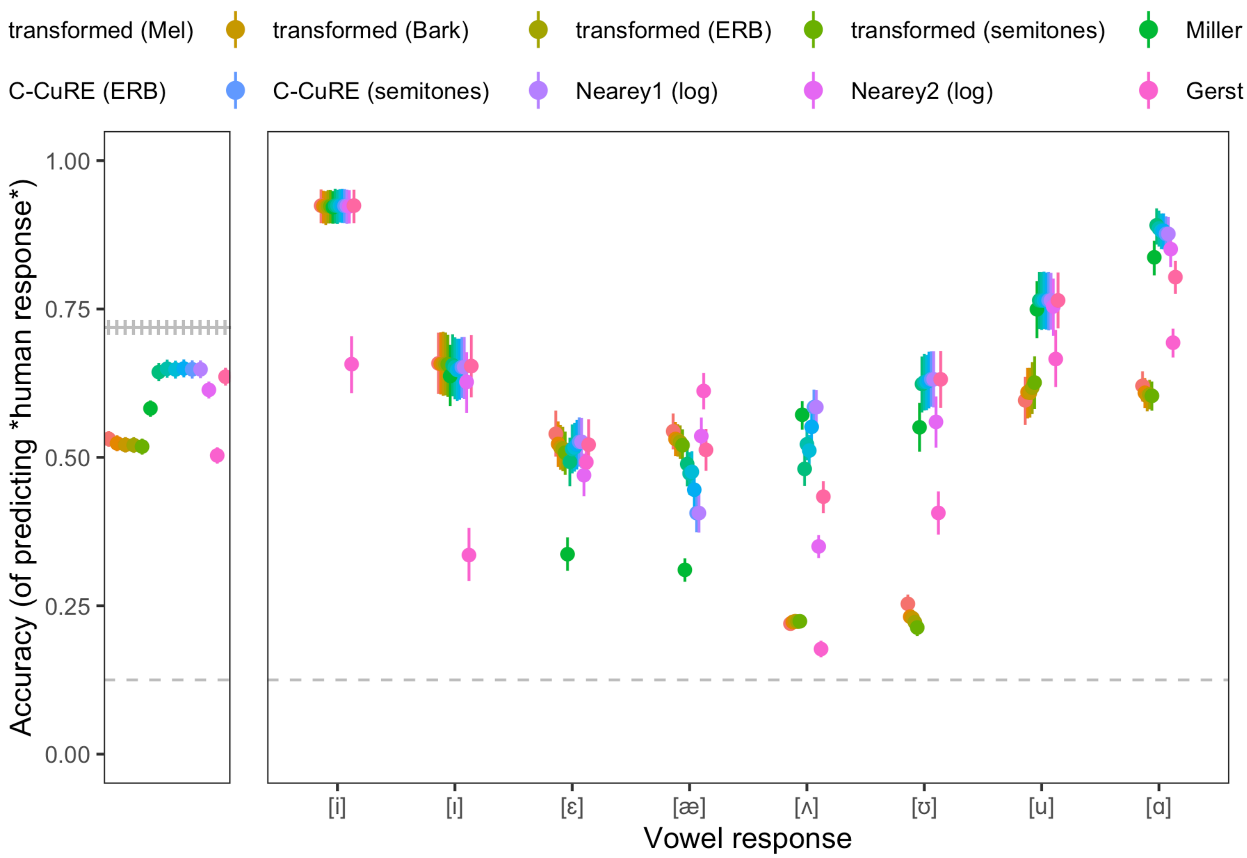
```
$n
NULL

$max_highlight
[1] 5

$unhighlighted_params
list()

$use_group_by
NULL

$use_direct_label
NULL

$line_label_type
[1] "ggrepel_label"

$label_key
<quosure>
expr: ^NULL
env:  empty

$label_params
$label_params$fill
[1] "white"


$keep_scales
[1] FALSE
```

```

$calculate_per_facet
[1] FALSE

attr(,"class")
[1] "gg_highlighter"
```

Response accuracy if human responses are to predict human responses. For this, we are assuming that someone knows the distribution of human responses for all items and uses the criterion choice rule to predict what the response would be. This gives us an upper limit—no model could ever predict human responses more accurately.

```
$predicates
<list_of<quosure>>

[[1]]
<quosure>
expr: ^IO.cue_normalization == "C-CuRE (Hz)"
env:  global


$n
NULL

$max_highlight
[1] 5

$unhighlighted_params
list()

$use_group_by
```

```
NULL

$use_direct_label
NULL

$line_label_type
[1] "ggrepel_label"

$label_key
<quosure>
expr: ^NULL
env:  empty

$label_params
$label_params$fill
[1] "white"


$keep_scales
[1] FALSE

$calculate_per_facet
[1] FALSE

attr(,"class")
[1] "gg_highlighter"
```



Fit log model to compare predictions of the ios (test of sign)+ evaluate against max accuracy and chance

6 Conclusions

In conclusion, these results indicate that pre-linguistic normalization (or computationally similar algorithms) contribute to the remarkable adaptive abilities of human speech perception. Our results further suggest that human perception only employs simple normalization operations—such as centering cues relative to a talker's mean. Finally, we find that simple ideal observers achieve performance far above chance in predicting perception, although our results also indicate that human perception might employ additional adaptive algorithms (given the 7.5% accuracy difference between the best performing models and human performance).

References

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*, 752–782.

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*, 148–203. https://doi.org/10.1037/a0038695

Kronrod, Y., Coppess, E., & Feldman, N. H. (2016). A unified model of categorical effects in consonant and vowel perception. *Psychological Bulletin and Review*, 1681–1712. https://doi.org/10.3758/s13423-016-1049-y