Comparing pre-linguistic normalization models against US English listeners' vowel perception

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

[1]Swedish Language and Multilingualism, Stockholm University, Sweden
[2]Brain and Cognitive Sciences, University of Rochester, USA

https://doi.org/10.36505/ExLing-2022/13/0037/000579

Abstract

We investigate the role of pre-linguistic normalization in the perception of US English vowels. We train Bayesian ideal observer (IO) models on unnormalized or normalized acoustic cues to vowel identity using a phonetic database of 8 /h-VOWEL-d/ words of US English. We then compare the IOs' predictions for vowel categorization against L1 US English listeners' 8-way categorization responses for recordings of /h-VOWEL-d/ words in a web-based experiment. Results indicate that pre-linguistic normalization substantially improves the fit to human responses from 74% to 90% of best-possible performance.

Keywords: speech perception, vowel normalization, computational model

Introduction

One of the central challenges for human speech perception is that talkers differ in pronunciation – i.e., how they map linguistic categories and meanings onto the acoustic signal. While this challenge is always present, it is most evident when listeners first encounter talkers with unfamiliar pronunciations. What mechanisms allow listeners to overcome this challenge – often rapidly, even after brief exposure – remains unclear.

One highly influential hypothesis holds that inter-talker differences are removed via low-level pre-linguistic auditory normalization of acoustic cues. There is now at least a dozen of competing normalization proposals (e.g., Lobanov, 1971; Nearey, 1978). Previous work has found that normalization reduces inter-talker variability due to, e.g. anatomical or physiological factors (e.g., Adank et al., 2004; Disner, 1980; Labov, 2010). This leaves open whether listeners actually employ normalization, and which normalization approach best explains listeners' vowel categorization. Only a relatively small number of studies has addressed these questions (e.g. Richter et al., 2017, for US English). Here, we contribute to this line of research by comparing normalization accounts against novel data on the perception of US English vowels.

Methods

Predicting speech perception from phonetic databases

To compare how well different normalization approaches explain listeners' vowel perception, we employ a model of Bayesian inference, ideal observers (IOs, see e.g., Kleinschmidt & Jaeger, 2015). To provide predictions about human perception, IOs need estimates of the (1) the prior probability of the vowels in the current context and (2) vowel-specific cue distributions.

Since we use the IOs to provide predictions for an 8-way forced choice categorization experiment (see below), we set (1) to a uniform prior of .125 for each of the eight vowels. We obtained (2) from a phonetically annotated database of L1 US English vowel productions (Xie & Jaeger, 2020). This assumes—as do all major theories of speech perception—that listeners acquire implicit knowledge of the category-specific distribution of phonetic cues. The database includes 1,240 recordings of eight VOWEL-d words (*heed, hid, head, had, odd, hut, hood, who'd*, N=9 tokens per word from each of 17 female and male talkers). All words are annotated for the first three formants (F1-F3) as well as the mean fundamental frequency (F0). IOs were trained on the unnormalized or normalized F1 and F2 cues, the primary cues to US English vowel identity.

Specifically, we considered nine types of normalization. The first four transform F1 and F2 from the untransformed *acoustic* space (Hz) into one of four *perceptual* spaces hypothesized to underlie human auditory perception (Mel, Bark, ERB, and semitones). The remaining five approaches constitute normalization in a narrower sense: they center and/or standardize F1 and F2 based on their marginal distribution *across* all eight vowels (e.g., Gerstman, 1968; Lobanov, 1971; Miller, 1989; and two approaches in Nearey, 1978).

While the Xie & Jaeger database is comparatively large for a phonetically annotated corpus, it is small compared to the amount of input that human learners receive during language acquisition. To avoid over-fitting IOs to the database, we used 5-fold cross-validation: we trained five different IOs for each of the 10 different unnormalized, transformed, or normalized approaches. Each IO was trained on 80% of the recordings from each vowel of each talker in the database. The predictions of each of the 5 * 10 IOs were then compared against human responses from a perception experiment described next.

Vowel categorization experiment

We exposed L1 US English listeners (N=22) to the h-VOWEL-d productions of one female L1 US English talker from the Xie & Jaeger database. The experiment was administrated on Amazon Mechanical Turk and consisted of 144 trials (9 recordings per vowel * 8 vowels * 2 repetitions). On each trial, participants saw all 8 h-VOWEL-d words displayed on screen (order counter-balanced across participants) and then heard one of the recordings (in randomized order, grouped by repetition of the recording into two blocks). Participants were instructed to click on the word they heard the talker say.

Results and discussion

Human performance

Participants' responses matched the vowel intended by the talker on 71.1% of all trials. This illustrates the challenge posed by cross-talker variability and individual differences in listeners' language backgrounds: without informative exposure to the unfamiliar talker and in the absence of disambiguating context, listeners categorize recordings incorrectly in at least 1 of 4 cases! We then calculated, for each of the 72 recordings, how much listeners agreed on its categorization. On average, the most frequent response for a recording was given on 72% of all trials (out of 2 trials for each of the 22 participants). This provides an important reference against which to compare model performance: 72% recognition accuracy is what one would achieve in predicting human performance if one employs the accuracy-maximizing decision rule (criterion choice), and always categorizes recordings based on the most frequent responses given by listeners (henceforth *expected ceiling performance*).

Model performance

The performance of the IOs was assessed by comparing their predictions for human responses, i.e., their posterior probability of inferring human categorization responses (Figure 1). We make five observations. (1) All models overall perform substantially above chance (Figure 1, left panel). (2) Transformations from the acoustic space into perceptual spaces does not improve model performance, but (3) normalization can: IOs trained on normalized cues, perform significantly better than the IO trained on unnormalized cues (53.1%, SE=0.3%, $p<2e\text{-}16$), except for Gerstman normalization (mean accuracy 50.3%, SE=0.5%). The two highest performing IOs employ Nearey's log-mean (mean accuracy 64.9%, SE=0.5%) or Lobanov normalization (mean accuracy 63.6%, SE=0.5%). The high performance of general standardizing procedures, such as Nearey and Lobanov, replicates previous findings, both from studies comparing against human responses (e.g., Richter et al., 2017) and simulated responses (e.g., Escudero & Bion, 2007). Gerstman normalization, however, still outperformed untransformed models in previous studies (unlike here). The improvements due to normalization are substantial: the unnormalized IO achieves 74% of the best possible performance (the expected ceiling performance), whereas the best performing normalization IOs achieve 90% of the best possible performance.

(4) Points 1-3 also hold for each vowel is separately (right panel). Finally, (5) no single normalization procedure outperforms all others normalization procedure on *all* vowels (right panel, Figure 1). Even for the two best-fitting IOs, there is at least one vowel for which they are not among the best models.

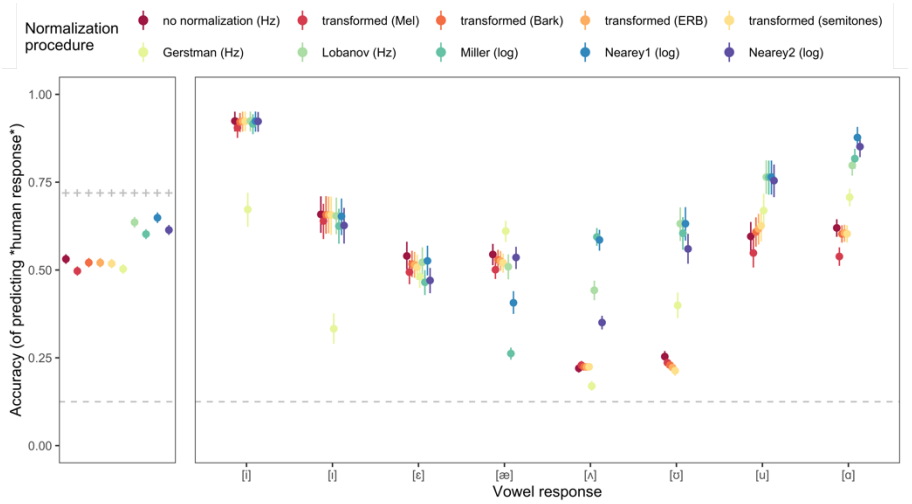

Figure 1. Prediction accuracy of 10 ideal observers for human vowel responses. **Left:** Overall accuracy across vowels. Plus line indicates expected ceiling performance (see text). **Right:** per-vowel accuracy. Dots indicate mean accuracy across the five folds. Intervals show average bootstrapped 95% confidence intervals across the five folds, thus indicating uncertainty about model's accuracy in predicting human performance. Grey line indicates chance.

Our results suggest that pre-linguistic normalization (or computationally similar algorithms) provide a plausible explanation for the remarkable adaptive abilities of human speech perception. We find that models based on normalized F1 and F2 cues can achieve up to 90% of the achievable accuracy. Future work should determine whether the remaining 10% can be achieved by adding additional cues (e.g., F3 or vowel duration), or whether they point to additional mechanisms (e.g., representational changes or changes in decision-making, Xie, Jaeger, & Kurumada, 2022). Similarly, it is possible that the effects we observed for normalization could be accounted for by alternative mechanisms (ibid).

References

Escudero, P., Bion, R.A.H. 2007. Modeling vowel normalization and sound perception as sequential processes. ICPhS 16. 1413-16.

Richter, C., Feldman, N.H., Salgado, H., Jansen, A. 2017. Evaluating low-level speech features against human perceptual data. ACL 5, 425-40.

Xie, X., Jaeger, T.F. 2020. Comparing non-native and native speech: Are L2 productions more variable? The Journal of the Acoustical Society of America 147. 3322-47.